MDPI

Article

A Comparative Study for Patch-level and Pixel-level Segmentation of Deep Learning Methods on Transparent Images of Environmental Microorganisms: from Convolutional Neural Networks to Visual Transformers

Hechen Yang [1], Xin Zhao [1], Tao Jiang [2]*, Jinghua Zhang [1], Peng Zhao [1], Ao Chen [1], Marcin Grzegorzek [3], Shouliang Qi [1], Yueyang Teng [1],Chen Li [1]*

[1] Microscopic Image and Medical Image Analysis Group, MBIE College, Northeastern University, 110169, Shenyang, PR China;
[2] School of Control Engineering, Chengdu University of Information Technology, Chengdu 610225, China;
[3] Institute of Medical Informatics, University of Luebeck, Luebeck, Germany;
* Correspondence: jiang@cuit.edu.cn(T.J.); lichen201096@hotmail.com.(C.L.)

Abstract: Nowadays, the field of transparent image analysis has gradually become a hot topic. However the traditional analysis methods are accompanied by a large number of carbon emissions, and this process consumes a lot of manpower and material resources. With the continuous development of computer vision, it is more appropriate to use computers to analyze images. However, the low contrast between the foreground and background of transparent images makes it difficult for computers to segment transparent objects. For this problem, we start the analysis with pixel patches in the image, and then classify the patches as foreground and background. Finally, the segmentation task of transparent images is completed through the reconstruction of pixel patches. In order to facilitate people to understand the performance of different deep learning networks for transparent image segmentation, this paper conducts a series of comparative experiments using patch-level and pixel-level methods. In two sets of experiments, we compared the segmentation performance of four *Convolutional Neural Network* (CNN) models and one *Visual Transformers* (ViT) model on the transparent Environmental Microorganism Data Set Fifth Version dataset, respectively. The research results show that U-Net++ has the highest accuracy rate in the pixel-level comparison experiment with a value of 95.32%. In the patch-level comparison experiment, the highest accuracy rate is ResNet50 with a value of 90.00%. Furthermore, ViT has the lowest accuracy of 89.25% on patch-level segmentation experiments. However, the accuracy rate of ViT in the pixel-level segmentation experiment is 95.31%, second only to U-Nnet++. We conclude that ViT performs the worst in segmentation experiments on pixel-level segmentation, but outperforms most convolutional neural networks on patch-level segmentation. This conclusion is also verifie by the *Environmental Microorganism Data Set Sixth Version dataset* (EMDS-6).

Keywords: Patch-Level; Pixel-Leve; Image Classification; Image Segmentation; Transparent Images; Deep Learning; Convolutional Neural Network; Visual Transformer;Environmental Microorganism

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

With the advent of the era of science and technology, the application of transparent images has become more and more widely used in various fields around humans, such as the segmentation of renal transparent cancer cell nuclei in medicine [1]. The shape and location information of the cell nucleus is of great significance for the segmentation and diagnosis of benign and malignant renal cancer [2] [3]. Another example is identifying the number of transparent microorganisms in the environment to judge the degree of environmental pollution [4]. In recent years, the segmentation of transparent objects in images is also a hot spot in vision research [5] [6]. It is not an easy task to detect

whether there are transparent objects or translucent objects in images [7]. Because the transparent object area to be observed is generally very small or thin, the colors and contrast of foreground and background are similar. Only the residual edge part leads to the low resolution of foreground or background, which largely depends on its background and lighting conditions. Therefore, there is an urgent need for effective methods to identify transparent or translucent images.

In recent years, deep Learning has good performances in the field of computer vision [8]. For example, in [9], a deep learning model is developed to detect and track sperm, which can effectively assist doctors in judging male reproductive health. In [10–13], a deep learning network is used to identify areas of cervical cancer to help doctors analyze cervical histopathology images. Due to the continuous increase of Corona Virus Disease 2019 (COVID-19), the workload of doctors' detection is also increasing. In [15], the detection performance of 15 different deep learning models for COVID-19 X-ray image identification are compared, which can help reduce the workload of doctors. In [16], a multiple network model is proposed for the analysis of intracranial pressure (ICP) and heart rate (HR) behavior after severe traumatic brain injury in pediatric patients. In [17], a deep learning model is developed to help pathologists detect cancer subtypes or genetic mutations. In [18], a deep learning model is trained on clinical data. This model achieves the prediction of response to immune checkpoint inhibitors in advanced melanoma, effectively assisting doctors in diagnosis. In [19], machine learning methods are used to realize the investigation, prediction and discrimination of COVID-19. We consider the excellent performance of computer vision in image analysis [20], such as high speed, high accuracy, low consumption, high degree of quantification, strong objectivity [21]. In addition, computer analysis of image is more energy-saving and emission-reducing than traditional methods, greatly reducing energy consumption. Therefore computer vision can make up the shortcomings of traditional morphological methods [22]. It brings new opportunities to transparent image analysis [23]. Especially when the object is transparent or has low contrast in the image, we need more foreground information, so we usually find more visual details to recover the lost information from patches or pixels. As shown in Fig 1, the foreground and background of microorganisms are similar. There is only a small amount of information on the edges, so it is difficult for traditional CNN algorithms to globally distinguish transparent objects in images. [24]. For this problem, it is necessary for us to analyze transparent images from patches. We crop the image into fixed-size patches and lead deep learning network to learn the features of the visual information of foreground and background patches. The network trained in this way is sensitive to the foreground and background, which helps to distinguish transparent objects and achieve the purpose of segmentation.

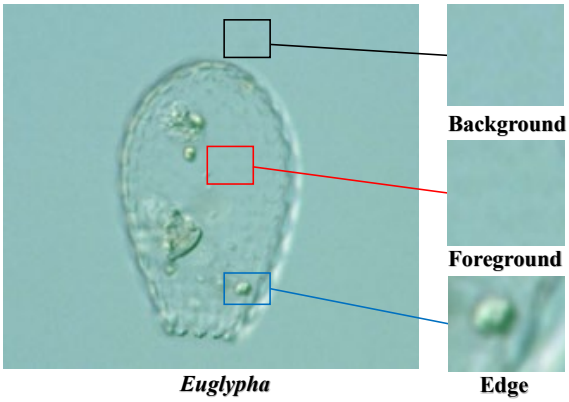


Figure 1. An example of transparent images (a low contrast environmental microorganism image).

In recent years, machine vision has been widely used in image processing [25] [26]. Deep learning is a more effective method in the field of machine vision, such as the popular *Convolutional Neural Network* (CNN) X-Inception [27], VGG-16 [28], Resnet50 [29], Inception-

V3 [30], U-Net [31], and novel *Visual Transformers* (ViT) [32]. CNNs gradually expand the receptive field by increasing the size of the convolution kernel until it covers the entire image, so CNNs complete the image extraction from local to global information. In contrast, transformers can obtain global information from the beginning, making learning more challenging, but their ability to retain long-term dependence is more potent than CNN [32]. Hence, CNNs and Transformers have advantages and disadvantages in dealing with visual information. Therefore, this paper compares patch-level and pixel-level segmentation performance of transparent images with different CNN and Visual Transformers methods. It aims to discover the adaptability of various deep learning models in this research domain.

The main contributions of this paper are as follows:

(1) A comparative study on patch-level transparent image segmentation is carried out to help people to analyze transparent images.

(2) The segmentation performance of multiple CNN and ViT deep learning networks under patch-level and pixel-level images are compared, which is convenient for people to do further ensemble learning.

2. Related Work

This section briefly introduces the related research on transparent images in practical analysis tasks and the classical deep learning models.

2.1. Introduction to Transparent Image Analysis

Object analysis is one of the essential branches in robot vision, especially the analysis of transparent images of objects (transparent images) is challenging [33]. In the traditional machine learning method, the multi-class fusion algorithm can only extract the shallow features of the transparent image, and the obtain feature layer is incomplete. In practical applications, it is difficult for multi-class fusion algorithms to detect transparent objects. For example, home robots can not see things at all when they are detecting some transparent glassware. The ClearGrasp machine learning algorithm performs well in analyzing transparent objects [34]. It can estimate high-precision data of transparent objects from RGB-D transparent images, thereby improving the accuracy of transparent object detection.

As a necessary technical means for analyzing objects, photoelectric sensors are widely used in industrial automation, mechanization, and intelligence. It uses the properties of light to detect the position and change of the object, but when detecting transparent color objects, the light beam of the traditional diffuse reflection photoelectric sensor penetrates the transparent material, causing the sensor to fail. Diffuse reflection photoelectric sensor adopts a phase-locked loop narrowband filter frequency selection technology, which improves the sensitivity to self-returning light and stability of detecting transparent objects [35].

There are many transparent objects in the industrial field, such as transparent plastics, transparent colloids, and liquid drops. These transparent objects bring much uncertainty to products. If factories want to have high-quality products, sometimes it is essential to analyze these transparent objects and control the shapes of the transparent objects. However, it is a difficult problem to segmentation the shape of transparent objects through morphological methods. For instance, Hata et al. use a genetic algorithm to segmentation the transparent paste drop shape in the industry and obtain good performance [36].

The segmentation of transparent objects is very useful in computer vision applications. However, the foreground of a transparent image is usually similar to its background environment, which leads to the general image segmentation methods in dealing with transparent images in general. The light field image segmentation method can accurately and automatically segment transparent images with a small depth of field difference and improve the accuracy of the segmentation, and it has a small amount of calculation [37]. Hence, it is widely used in the segmentation of transparent images.

The correct segmentation of zebrafish in biology has extensively promote the development of life sciences. However, the zebrafish's transparency makes the edges blurre in the

segmentation. The mean shift algorithm can enhance the color representation in the image and improve the discrimination of the specimen against the background [38]. This method improves the efficiency and accuracy of zebrafish specimen segmentation.

Visual object analyze is vital for robotics and computer vision applications. Commonly use statistical analyze methods such as bag-of-features [39] are often applied to image segmentation. The principle is to extract local features of the image for segmentation. However, the foreground transparent objects in transparent images do not have complete features, so these methods are difficult to accurately segment transparent images. The more popular method is the light field distortion feature [40], which can describe transparent objects without knowing the texture of the scene, thus improving the accuracy of segmentation transparent images.

2.2. Introduction Classic Deep Learning Network Models

Simonyan et al. propose the VGG series of deep learning network models (VGG-Net), of which VGG-16 is the most representative [41]. VGG-Net can imitate a larger receptive field by using multiple 3×3 filters, enhancing nonlinear mapping, reducing parameters, and improving the network to be more judgmental. Meanwhile, VGG-16 continues to deepen the previous VGG-Net, with 13 convolutional layers and three fully connected layers. With the continuous increase of convolution kernel and convolution layer, the nonlinear ability of the model is stronger. VGG-16 can better learn the features in images and achieve good performance in analyzing image classification, segmentation, and detection. Simonyan proves that as the depth of the network increases, it promotes the accuracy of image analysis [41]. Nevertheless, this increase in depth is not without a limit. Excessively increasing the depth of the network will lead to network degradation problems. Therefore, the optimal network depth of VGG-Net is set to 16-19 layers. Moreover, VGG-16 has three fully connected layers, which causes more memory to be occupied, too long training time, and difficulty in tuning parameters.

He et al. propose the ResNet series of networks and add a residual structure in networks to solve the problem of network degradation [42]. The ResNet model introduces a jumpy connection method "shortcut connection". This connection method allows the residual structure to skip some levels that not be fully train in the feature extraction process and increases the model's utilization of feature information during the training process. As the most classical model in the ResNet series, ResNet50 has a 50-layer network structure. This model adopts the highway network structure, which makes the network have strong expression capabilities and acquire more advanced features. Therefore, it is widely used in the field of image analysis. However, the network model is too deep and complicated, so how to judge which layers in the deep network not be thoroughly train and then optimize the network is a complex problem.

Szegedy et al. propose the GoogLeNet network model, which has the advantage of reducing the complexity of the network based on ResNet. They first propose Inception-V1, whose network is 22 layers deep and consists of multiple Inception structures cascade as basic modules. Each Inception module consists of a 1×1, 3×3, 5×5 convolution kernel and a 3×3 maximum pooling, which is similar to the idea of multi-scale and increases the adaptability of the network to different scales [43]. With the continuous improvement of the Inception module, the Inception-V2 network uses two 3×3 convolutions instead of 5×5 convolutions. It increases the BN method, which reduces the amount of calculation and speeds up the training time [44]. The Inception-V3 network introduces the idea of decomposing convolution, splitting a larger two-dimensional convolution into two smaller one-dimensional convolutions, further reducing the amount of calculation [45]. At the same time, Inception-V3 optimizes the Inception module embeds the branch in the branch and improves the model's accuracy.

X-Inception is another improvement after Inception-V3 [46]. It mainly uses depth-wise separable convolution to replace the convolution operation in Inception-V3. The X-Inception model uses deep separable convolution to increase the width of the network,

which improves the accuracy of the classification and improves the ability to learn subtle features. Meanwhile, X-Inception adds a residual mechanism similar to ResNet to significantly improve the speed of convergence during training and the model's accuracy. However, X-Inception is relatively fragmented in the calculation process, which results in a slower iteration speed during training.

U-Net is a convolutional neural network, which is initially used to perform the task of medical image segmentation. The architecture of U-Net is symmetrical. It consists of a contracting path and an expansive path [31]. There are two significant contributions of U-Net. The first is the strong use of data augmentation to solve the problem of insufficient training data. The second is its end-to-end structure, which can help the network retrieve the information from the shallow layers. With outstanding performance, U-Net is widely used in semantic segmentation.

Transformer is a deep neural network based on the self-attention mechanism, enabling the model to be trained in parallel and obtain the global information of the training data. Due to its computational efficiency and scalability is widely used in Natural Language Processing. Recently, Dosovitskiy et al. proposed the Vision Transformer (ViT) model and find that it performs very well on image classification tasks [47]. In the first step of training, the ViT model divides pictures into fixed-size image patches and uses its linear sequence as the input of the transformer model. In the second step, position embeddings are added to the embeddings patches to retain the position information, and then the image features are extracted through the multi-head attention mechanism. Finally, the classification model is trained. ViT breaks through the limitation that CNN model cannot be calculated in parallel, and self-attention can produce a more interpretable model. ViT is suitable for solving image processing tasks, but experiments have proved that large data samples are needed to improve the training effect.

3. Comparative Experiment

This section introduces patch-level and pixel-level segmentation experiments and segmentation results of transparent images under several deep learning networks. The workflow of patch-level and pixel-level image segmentation is shown in Fig. 2.

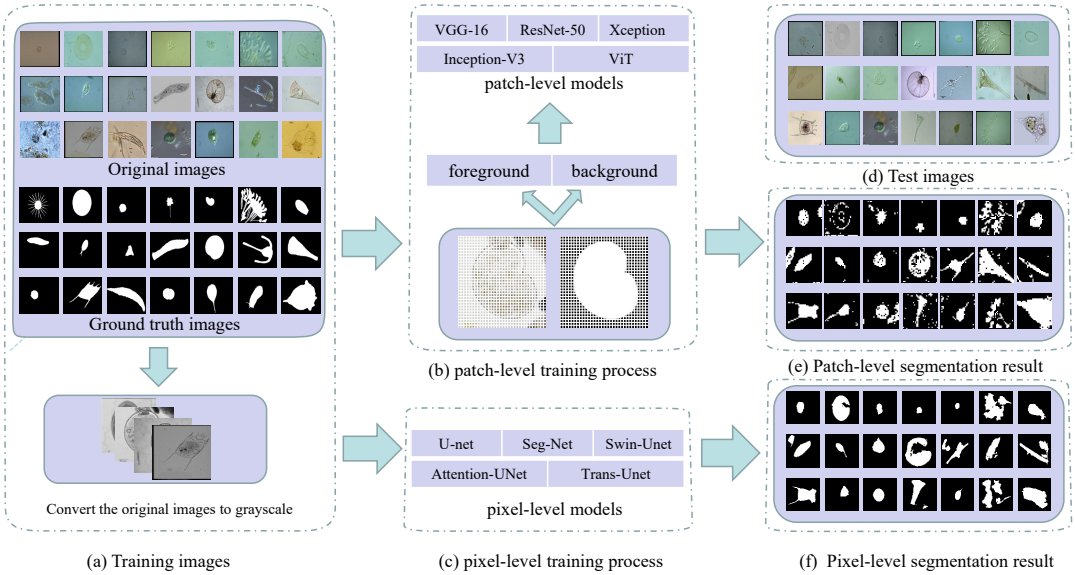


Figure 2. Workflow of patch-level and pixel-level segmentation in transparent images (using environmental microorganism EMDS-5 images as examples) ((a) is the image of the training set and the grayscale of the original image. (b) is the patch-level and pixel-level training process. In (d) is the test set image. (e) and (f) are patch-level and pixel-level segmentation results, respectively).

3.1. Experiment Setting

3.1.1. Data Settings

In our work, we use Environmental Microorganism Data Set Fifth Version (EMDS-5) as transparent images for analysis [4]. Tab 1 shows the data distribution of EMDS-5 in the experiment. It is a newly released version of the EMDS series, which includes 21 types of EMs, each of which contains 20 original microscopic images and their corresponding ground truth (GT) images (examples are shown in Fig.3). We randomly divide each category of EMDS-5 into training, validation, and test data sets at a ratio of 1:1:2. Therefore, we have 105 original images and their corresponding GT images for training and validation, respectively, and 210 original images for testing.

Table 1. EMDS-5 Experimental data.

	Training Set	Validation Set	Test Set
Actinophrys	5	5	10
Arcella	5	5	10
Aspidisca	5	5	10
Codosiga	5	5	10
Colpoda	5	5	10
Epistylis	5	5	10
Euglypha	5	5	10
Paramecium	5	5	10
Rotifera	5	5	10
Vorticlla	5	5	10
Noctiluca	5	5	10
Ceratium	5	5	10
Stentor	5	5	10
Siprostomum	5	5	10
K.Quadrala	5	5	10
Euglena	5	5	10
Gymnodinium	5	5	10
Gonyaulax	5	5	10
Phacus	5	5	10
Stylonychia	5	5	10
Synchaeta	5	5	10
total	105	105	210

3.1.2. Data Preprocessing

Patch-Level Data Preprocessing:

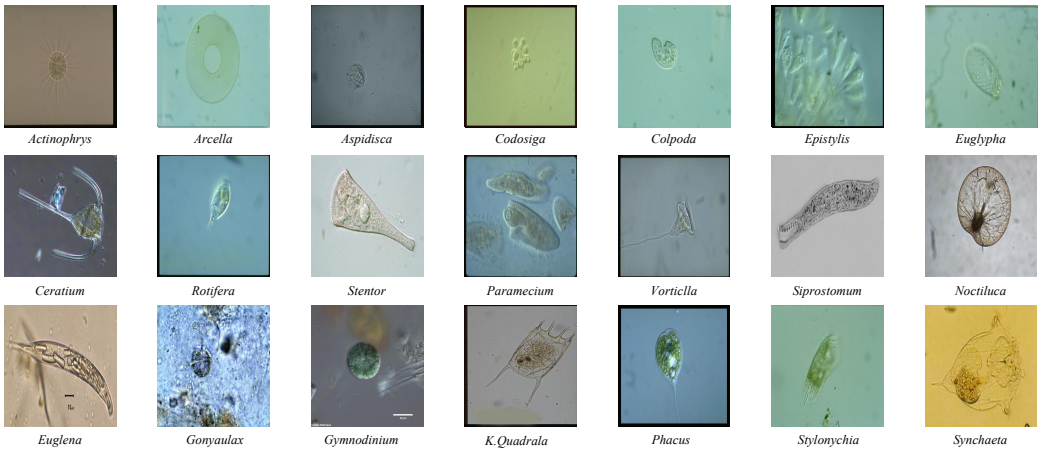
In the first step, considering the colour information is inefficient in EM segmentation [48], these image are converted into grayscale. In the second step, we convert all the image sizes into 256×256 pixels uniformly due to the microscopic images having various sizes. In the third step, the training and validation images and their corresponding GT images are cropped into patches (8×8 pixels), where 105×1024=107520 patches are obtained. We divide these small patches into two categories according to the corresponding GT image small patches: foreground and background. The partition criterion is whether the interest area in the patch takes up half of the whole patch. If it is, we will assign foreground as the label of this patch; otherwise, it is annotated the background. Last step, we find that the 8×8 pixels patches with foreground and background are 16554 and 90966, respectively. During the training process, we find that the model weights are heavily biased towards negative samples due to the imbalance of positive and negative samples. In order to avoid data imbalance during training, we rotate the training set image small patches by 0, 90, 180, 270 degrees and mirror them for data augmentation. Then we further obtain 16554×8=132432 patches, from which 90966 patches are randomly selected as the finally used patches in the training set. The details of the image patches are shown in Tab. 2.

Table 2. Patch-Level Data Preprocessing. FG (foreground) and BG (background)

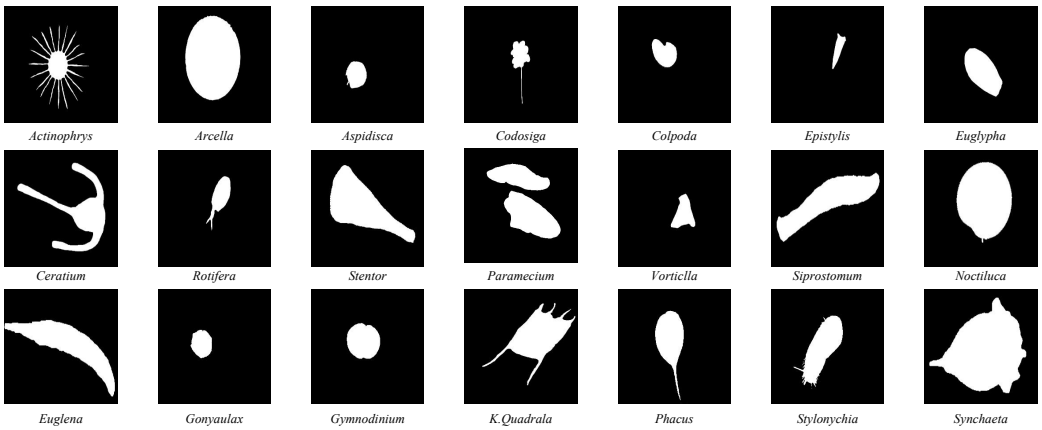
Data Set	Training Set	Validation Set	test Set
8 × 8 pixels FG	16554	17356	32445
8 × 8 pixels BG	90966	90164	182595
Augmentation With FG	90966	\	\
8 × 8 Total	181932	107520	215040

Pixel-Level Data Preprocessing:

We convert the image to grayscale and resize the image size to 256 × 256 pixel for the pixel-level segmentation experiments.



(a) Original Image



(b) Ground truth image

Figure 3. Examples of the environmental microorganism image in EMDS-5. (a) is the original images of EMDS-5, each image contains one or more EM objects of the same species, and one image is selected for each species as a representative. (b) correspond to the real segmentation images of microorganisms in each image in the (a). The pixel value of the background part in the microorganism image is set to 1, and the foreground part is set to 0.)

3.1.3. Experimental Environment

This comparative experiment is conducted on a local computer. The running memory of the computer is 16 GB. The computer uses Win10 Professional operating system, and it is equipped with an 8 GB NVIDIA Quadro RTX 4000 GPU. In the patch-level experiment, The four CNN network models are imported from keras version 2.3.1 and use tensorflow 2.0.0 as the background. The experimental frameworks for ViT and pixel-level are Pytorch 1.7.1 and Torchvision 8.0.2.

3.1.4. Hyper Parameters

The patch-level experiment uses the Adam optimizer with a 0.0002 learning rate and sets the batch size to 32. In Fig. 4 show the accuracy and loss curves of different deep learning models in this experiment. The Epoch is determined according to the convergence of the loss curve. In our pre-test, we tried to train 100 epochs and keep the best training model weights, and we found that the best model appear between 40 and 50, where too much training caused overfitting and too little training were not able to train the optimal model. Therefore, considering the computational performance of the workstation, we finally set 50 epochs for training. Meanwhile, because of the outstanding classification ability of CNN in ImageNet and the significant performance of transfer learning with limited training data set [41], we use the limited EM training data to fine-tune the CNN model pretrained by ImageNet [49] [50]. It is proved that the use of CNN pretrained on ImageNet is useful for classification tasks through the concept of transfer learning and fine-tuning in [51]. Before fine-tuning the pretrained CNN, we freeze the parameters of the pretrained model. After that, we use the patch-level data to fine-tune the dense layers of CNN. We keep the backbone network of the CNN classification network to extract image features, and replace the last fully connected layer of CNN model with Global Average Pooling2D + dense + dense + softmax. Global Average Pooling2D simplifies a large number of parameter operations. The purpose of the dense layer is to extract the correlation between these features through nonlinear changes in the dense layer, and finally map them to the output space. Finally, the class probability result is output through softmax. Meanwhile, we compare the validation set accuracy of the ViT model with and without pretrained weights. In both sets of experiments, we train three times and then take the average. We find that ViT without pretrained weights and ViT with imagenet pretrained weights had an accuracy of 0.8923 and 0.8926 on the validation set, respectively. During training, ViT takes about 2G less memory than loading the imagenet pre-trained weight model. To compare the performance of the two, we use ViT without pretraining as the optimization option. We set the network depth to 6, heads to 16, mlp_dim to 3000, dropout and emb_dropout to 0.1. The pixel-level experiment uses the Adam optimizer with a 0.001 learning rate and sets the batch size to 4. Fig. 5 show the loss curves of different deep learning models in this experiment. We find that the training curves began to converge after 90 epochs of iterations for the five models. To prevent overfitting, we finally set 100 epochs for training.

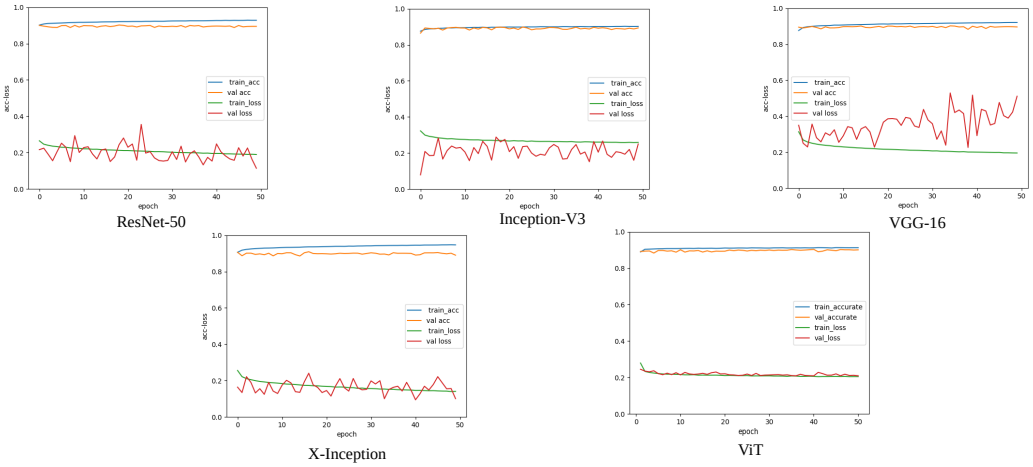


Figure 4. A comparison of the image segmentation results of the loss and accuracy curves of deep learning on 8×8 pixels training and validation sets.(Each legend has four curves, respectively, the accuracy and loss values of the training set, and the accuracy and loss values of the validation set.)

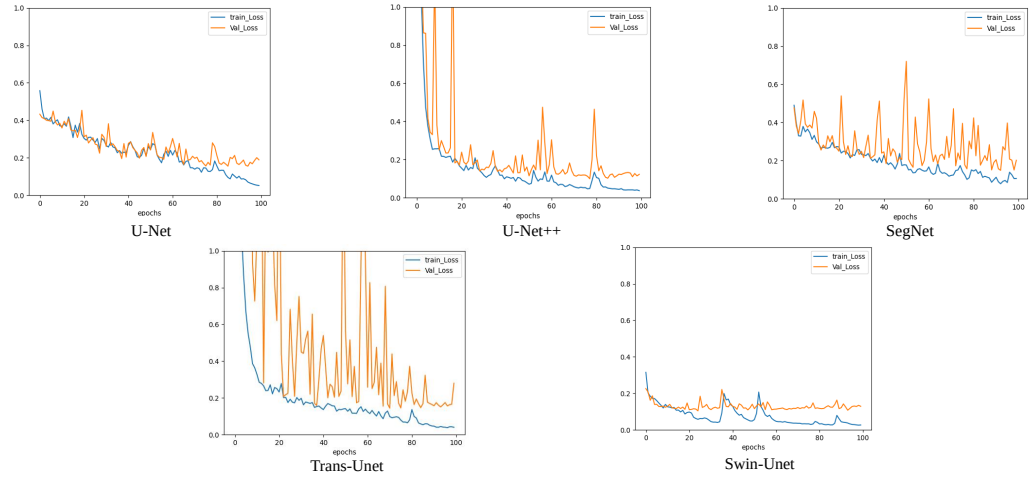


Figure 5. A comparison of the image segmentation results of the loss curves of deep learning on pixel-level training and validation sets.

3.2. Evaluation Metrics

To compare the classification foreground and background performance of different methods, we used the commonly used deep learning classification indexes Accuracy (Acc), Precision (Pre), Recall (Rec), Specificity (Spe), and F1-Score (F1) to evaluate the patch-level results [52]. Acc reflects the ratio of correct classification samples to total samples. Pre reflects the proportion of correctly predicted positive samples in model classification's positive samples. Rec reflects the correct proportion of model classification in whole positive samples. Spe reflects the proportion of the model correctly classifying the negative samples in the total negative samples. F1 is a calculation result that comprehensively considers the Pre and Rec of the model. Besides that, we employ Dice, Jaccard, Pre, Acc, and Rec to evaluate the results in pixel-level segmentation [16]. V_{pred} represents the foreground that the model predicts. V_{gt} represents the foreground in a ground truth image. From Tab 3, we can find that the higher the values of the first four metrics (Dice, Jaccard, Recall, and Accuracy) are, the better the segmentation results are. TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative) are concepts in the confusion matrix.

Table 3. Evaluation metrics.

Metrics	Formula	Metrics	Formula
Acc	$\frac{TP+TN}{TP+TN+FP+FN}$	Dice	$\frac{2\times\lvert V_{pred}\cap V_{gt}\rvert}{\lvert V_{pred}\rvert+\lvert V_{gt}\rvert}$
Pre (P)	$\frac{TP}{TP+FP}$	Jaccard	$\frac{\lvert V_{pred}\cap V_{gt}\rvert}{\lvert V_{pred}\cup V_{gt}\rvert}$
Rec (R)	$\frac{TP}{TP+FN}$	F1	$\frac{2\times P\times R}{P+R}$
Spe	$\frac{TN}{TN+FP}$		

3.3. Comparative Experiment

To avoid network model generalization, we perform five-fold cross-validation in all experiments in this paper. We take the average of the experimentally obtained model performance indicators as the data for the final evaluation model (Precision, Recall, F1-Score, Accuracy, Time, Size, Dice, Jaccard).

3.3.1. Comparative Experiment of Patch-level Segmentation

Comparison on Training and Validation Sets:

In order to compare the classification performance of CNNs and ViT models, we calculate precision, recall, F1-Score, and max accuracy are used to evaluate the models. The 5-class classification results of 8×8 pixels patches on validation set are presented in Tab 4. Overall, the Pre of the deep learning network classifying the transparent image background is higher than the foreground. Besides, the Pre of the five models to classify transparent images backgrounds is almost 97%; the highest is the VGG-16 value of 97.6%, and the lowest is the X-Inception and the ViT value of 96.7%. Meanwhile, the Pre rate of classification foreground VGG-16 is the best, and the Pre rate is 63.1%. The Inception-V3 obtains the lowest 53.3%. For transparent images foreground classification, the highest Rec rate is obtained with X-Inception, which is 89.2%, and the lowest one is ViT, which is 84.1%.For transparent images background classification, the highest Rec rate is the Vit value of 90.3%, and the lowest is the X-Inception value of 85.0%. The Spe obtained by the five models in the classification background is opposite to the Rec rate obtained in the classification foreground. Among the five models, the highest Acc is ResNet50 with a value of 92.87%, and the lowest is ViT with 89.26%.

Table 4. Classification performance of models of five-fold cross-validation experiment on validation set of 8×8 pixels patches. MAcc (Max Acc), FG (foreground) and BG (background) (In [%].)

Model	Class	Avg.Pre	Avg.Rec	Avg.Spe	Avg.F1	MAcc
ResNet50	FG	62.3	88.2	89.7	73.0	92.87
	BG	97.5	89.7	88.2	93.4	
Inception-V3	FG	61.8	88.6	89.5	72.8	90.24
	BG	97.6	89.5	88.6	93.4	
VGG-16	FG	63.1	88.6	90.0	73.7	92.09
	BG	97.6	90.0	88.6	93.6	
X-Inception	FG	53.3	89.2	85.0	66.7	91.10
	BG	96.7	85.0	89.2	90.9	
ViT	FG	62.4	84.1	90.3	71.6	89.26
	BG	96.7	90.3	84.1	93.4	

Comparison on Test Set:

In Tab 5 we summarize the results of these five network predictions. We can find that Acc of ResNet50 is the highest (90.00%), Acc of X-Inception is the lowest at 85.85%. Furthermore, the lowest prediction Acc of the transparent foreground is the X-Inception value of 51.8%, and the highest is the ResNet50 value of 62.2%.

In order to more intuitively express the classification results of CNN and ViT models for transparent image patches, we summarize the confusion matrices predicted by

Table 5. Classification performance of models of five-fold cross-validation experiment on test set of 8×8 pixels patches. PAcc (prediction accuracy), FG (foreground) and BG (background)(In [%].)

Model	Class	Avg.Pre	Avg.Rec	Avg.Spe	Avg.F1	Avg.PAcc
ResNet50	FG	62.2	87.2	90.6	72.6	90.0
	BG	97.5	90.6	87.2	93.9	
Inception-V3	FG	52.6	91.5	85.4	66.8	86.29
	BG	98.3	85.4	91.5	91.4	
VGG-16	FG	60.7	89.4	89.7	72.6	89.6
	BG	97.9	89.7	89.4	93.6	
X-Inception	FG	51.8	90.7	85.0	65.9	85.85
	BG	98.1	85.0	90.7	91.1	
ViT	FG	60.4	83.8	90.2	70.2	89.25
	BG	96.9	90.2	83.8	93.4	

five models into Fig. 6. We find that the ability of CNNs to classify foreground patches of transparent images is higher than that of ViT. Among them, the best CNN model is Inception-V3, which correctly classify 29686 foreground patches, accounting for 91.50% of the total correct foreground patches. ViT correctly classify 27177 foreground patches, accounting for 83.76% of the total correct foreground patches. In addition, the number of correctly classify backgrounds in ResNet50 is at most 165369, accounting for 90.57% of the total correct background patches, and the Pre of the classify background patches is 97.55%. Among the five models, ResNet50 has the highest prediction accuracy rate of 90.06%. The classification accuracy of X-Inception and Inception models is lower among the five models at 85.85% and 86.30%. Moreover, we find that the X-Inception and Inception models have poor background recognition performance, but better foreground recognition performance. The Inception-V3 model correctly classified up to 29,688 foreground patches, accounting for 91.50% of the total foreground patches. The X-Inception model misclassified a maximum of 27,409 background patches, accounting for 15.01% of the total background patches. The classification performance of the VGG model is relatively moderate among the five models. To better show the classification results, we reconstruct the transparent image after dicing in Fig. 7.

In Tab 6, we provide the model training and prediction time and the size of the model during the experiment. From the perspective of model training time, the ViT model is much lower than CNN models, where the ViT training time is 13992 seconds, and the X-Inception training time is the longest, 46383 seconds. From the perspective of the model's size, the minimum size of the ViT model is 31.2M, and the maximum size of the ResNet50 model is 114M. We calculate the time of the five prediction models. The fastest prediction time of Inception-V3 is 583 seconds, and the prediction time of a single picture is 0.0027 seconds. The slowest time of ViT is 1308 seconds, and the prediction time of a single image is 0.0061 seconds.

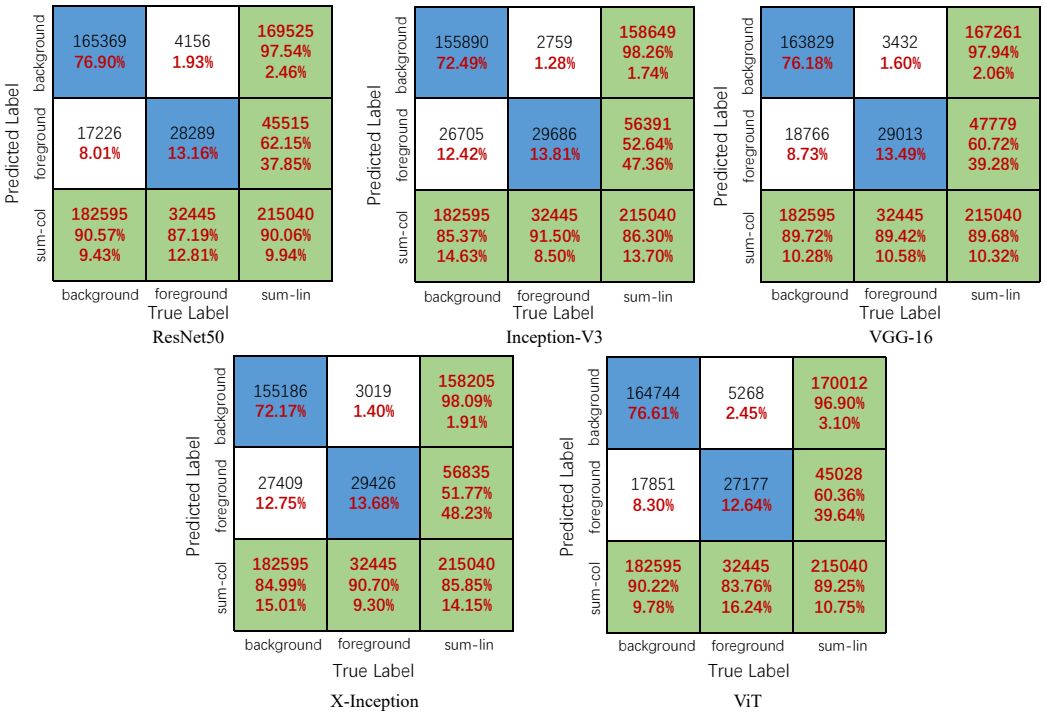


Figure 6. Predict the confusion matrix on test set of 8 × 8 pixels patches

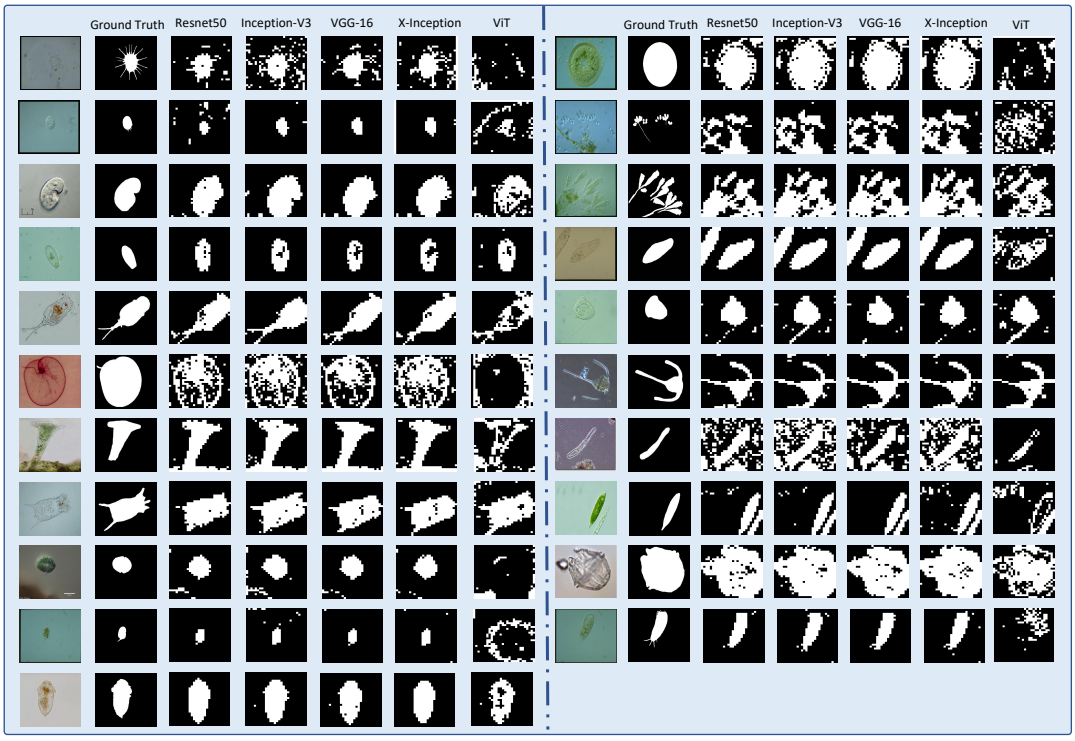


Figure 7. Reconstruct the 8 × 8 pixel patch transparent image segmentation results. (The figure contains the original image, ground truth image and Resnet50, Inception-V3, VGG-16, X-Inception, ViT network model predicted segmentation results.)

Table 6. A comparison of the classification results of five-fold cross-validationexperoment on train and test sets of 8 × 8 pixels patches. Train (Average training time), Test (Average test times) and Avg.p (Single picture prediction time)(In [s].)

model	Train	Test	Avg.p	Size(MB)
ResNet50	36754	878	0.0041	114
Inception-V3	24064	583	0.0027	107
VGG-16	34736	781	0.0036	62.2
X-Inception	46383	1014	0.0047	103
ViT	13992	1308	0.0061	31.2

3.3.2. Comparison Experiment of Pixel-Level Segmentation

To compare the effect of path-level segmentation, we conduct extended experiments on pixel-level segmentation. We apply five networks for comparative experiments: U-Net, U-Net++, SegNet, TransUnet, and Swin-UNet. We use these five networks to compare the performance of CNN and *visual transformer* (VT) for pixel-level segmentation. U-Net, U-Net++, SegNet stands for CNN network, Swin-UNet stands for transformer networks, TransUnet stands for CNNs joins transformer. In tab 7, we show five model prediction outcome metrics. We find that U-Net++ has the highest segmentation performance in the whole, but it also has the longest training time. U-Net has the worst segmentation performance. The segmentation result of vision transformer network (Swin-UNet) is second only to U-Net++. Its Jaccard and precision values of 71.26% and 85.00%, which are higher than other network models. In order to compare the segmentation results more intuitively, we show the pixel-level segmentation results in Fig. 8. We can see that pixel-level segmentation results are generally better than patch-level. However, the patch-level segmentation effect is better on multi-object transparent microorganism images. Compared with the 8 × 8 patch-level segmentation, the network model with transformer structure(Swin-UNet) at the pixel level performs well, and the VT is higher than the accuracy of the CNN network model. However, in the 8 × 8 patch-level experiment, the accuracy of CNN networks are higher than ViT. In Fig. 5, we find that the Loss curve stability of Swin-UNet is significantly better than the other four models during training. The stability of the training loss of the VT model is better than that of the CNN. In order to more intuitively reflect the training process of the model, we show the *Intersection-over-Union* (IOU) curves of the five models on the training set and the validation set in the pixel-level experiment in Fig. 9.

Table 7. Segmentation performance of models of five-fold cross-validation experiment on the test set

model	Avg.Dice	Avg.Jaccard	Avg.Precision	Avg.Recall	Avg.Acc
U-Net	71.82	59.23	68.98	76.06	91.93
U-Net++	82.51	73.51	83.42	85.98	95.32
SegNet	78.21	67.70	77.45	84.66	74.06
Trans-Unet	75.50	64.13	72.52	86.75	93.44
Swin-UNet	81.00	71.26	85.00	82.08	95.31

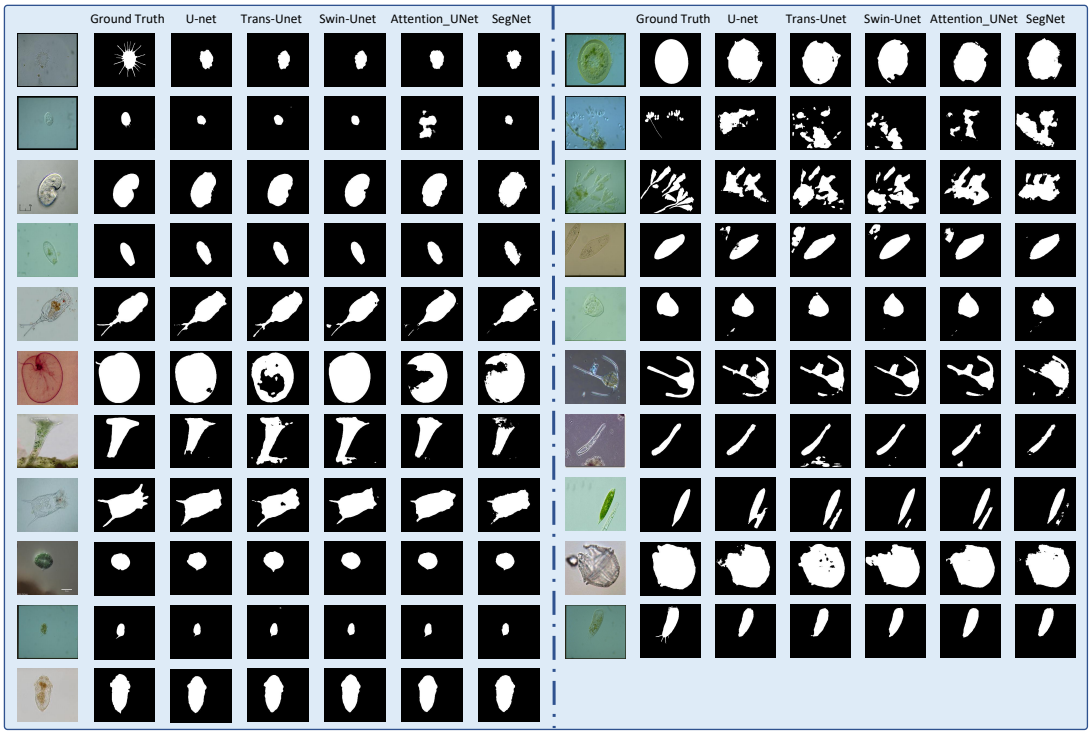


Figure 8. Reconstruction of pixel-level segmentation results on transparent images of the test set. (The figure contains the original image, ground truth image and U-net, Trans-Unet, Swin-Unet, Attention_Unet, SegNet. network model predicted segmentation results.)

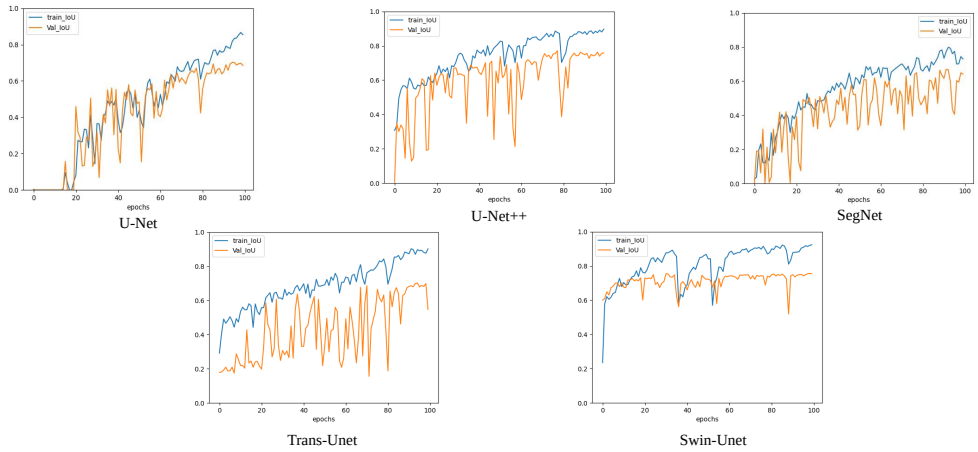


Figure 9. A comparison of the image segmentation results of the IOU curves of deep learning on pixel-level training and validation sets.

3.3.3. Additional experiment based on EMDS-6 dataset

To demonstrate the applicability of the models in our comparative experiments, we further compare five models in the pixel-level experiments on the EMDS-6 dataset [52]. The performance metrics of the experimental results of the five models are presented in Tab 8. We find that on EMDS-6, the pixel-level segmentation performance is basically consistent with the segmentation performance on EMDS-5. The segmentation performance of U-Net, U-Net++ and Swin-Unet models is similar, and the segmentation performance of segnet is the worst. In addition, the number of images in EMDS-6 is twice that of EMDS-5, so the model will learn more EMs information during training, which leads to an overall improvement in the segmentation performance of the five models.

Table 8. Segmentation performance of models of five-fold cross-validation experiment on the EMDS-6 test set.

model	Dice	Jaccard	Precision	Recall	Acc
U-Net	84.81	76.24	88.83	83.53	95.43
U-Net++	86.48	78.25	89.02	87.08	95.80
SegNet	74.63	62.50	73.88	83.59	91.21
Trans-Unet	84.66	76.087	86.04	86.88	94.98
Swin-UNet	86.11	78.05	89.46	85.79	95.49

3.4. In-depth Analysis

In the predicted 215040 patches, we compare the performance of five types of network classification foreground and background. In Fig. 6, we find that Inception-v3 has the largest number of correct foregrounds under 8×8 pixels patches. ResNet50 has the largest number of correctly classify backgrounds. We find that Inception-v3 has the largest number of correctly classify foregrounds under 224×224 pixels patches, and the largest number of correctly classify background patches is ViT. In addition, the number of foreground patches misclassify by the ViT network model is much smaller than that of the CNNs network. At the same time, the number of correctly classify foregrounds in the CNNs network is greater than that of the ViT network.

In the predicted 215040 patches, we compare the performance of five types of network classification foreground and background. In Fig. 6, we find that VGG-16 has the largest number of correct foregrounds under 8×8 pixels patches. Inception-v3 has the largest number of correctly classify backgrounds. However, the number of correctly classified foregrounds of ViT is higher than that of VGG-16, Inception-V3 and X-Inception. Besides, the ability of Swin-UNet to segment foreground also outperforms most models. So, VT model is also outstanding for low-transparency image recognition.

4. Conclusion and Future Work

In this paper, we aim at the problem that transparent images are difficult to segmentation by cropping the image into patches and classifying the foreground and background. We use CNNs and VT deep learning methods to compare patch-level and pixel-level performance of the segmentation of transparent images. We find that pixel-level generally outperforms patch-level in segmenting transparent microorganism images. However patch-level works better in multi-object segmentation. In addition, in the patch-level segmentation experiment, CNNs are better than the VT model, but in the pixel-level experiment, the VT model segmentation performance is better than most CNNs. When the patch pixel is smaller, the more regions perceived by the VT model, the stronger the ability to combine contextual information. In addition, the loss convergence and stability of the VT model during training are better than the CNN model. The VT model has great potential in the future. Therefore, CNN and ViT models have more advantages in image classification. CNN is good at extracting local features of images, while ViT is good at extracting global features of images combined with contextual information.

In the future, we plan to increase the amount of data to improve the stability of the comparison. Meanwhile, the images reconstructed by deep learning classification can be extended to the positioning, recognition, and detection of transparent images. We will further strengthen the application of results.

Author Contributions: Conceptualization, C.L.; methodology, H.Y. and C.L.; software, H.Y.; validation, P.Z., A.C. and H.Y.; formal analysis, H.Y.; investigation, M.G. and T.J.; resources, X.Z.; data curation, C.L. and H.Y.; writing—original draft preparation, H.Y. and C.L.; writing—review and editing, C.L., J.Z. and H.Y.; visualization, H.Y.; supervision, C.L.; project administration, C.L.; funding acquisition, C.L. and T.J. All authors have read and agreed to the published version of the manuscript.

Funding: National Natural Science Foundation of China (No. 61806047).

Acknowledgments: We thank Miss Zixian Li and Mr. Guoxian Li for their important discussion.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. S.-Y. Liao, O. N. Aurelio, K. Jan, J. Zavada, and E. J. Stanbridge, "Identification of the mn/ca9 protein as a reliable diagnostic biomarker of clear cell carcinoma of the kidney," *Cancer research*, vol. 57, no. 14, pp. 2827–2831, 1997.
2. D. Xue, X. Zhou, C. Li, Y. Yao, M. M. Rahaman, J. Zhang, H. Chen, J. Zhang, S. Qi, and H. Sun, "An application of transfer learning and ensemble learning techniques for cervical histopathology image classification," *IEEE Access*, vol. 8, pp. 104 603–104 618, 2020.
3. X. Zhou, C. Li, M. M. Rahaman, Y. Yao, S. Ai, C. Sun, Q. Wang, Y. Zhang, M. Li, X. Li *et al.*, "A comprehensive review for breast histopathology image analysis using classical and deep neural networks," *IEEE Access*, vol. 8, pp. 90 931–90 956, 2020.
4. Z. Li, C. Li, Y. Yao, J. Zhang, M. M. Rahaman, H. Xu, F. Kulwa, B. Lu, X. Zhu, and T. Jiang, "Emds-5: Environmental microorganism image dataset fifth version for multiple image analysis tasks," *Plos one*, vol. 16, no. 5, p. e0250631, 2021.
5. J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, and H. Li, "Lcu-net: A novel low-cost u-net for environmental microorganism image segmentation," *Pattern Recognition*, vol. 115, p. 107885, 2021.
6. F. Kulwa, C. Li, X. Zhao, B. Cai, N. Xu, S. Qi, S. Chen, and Y. Teng, "A state-of-the-art survey for microorganism image segmentation methods and future potential," *IEEE Access*, vol. 7, pp. 100 243–100 269, 2019.
7. M. P. Khaing and M. Masayuki, "Transparent object detection using convolutional neural network," in *International Conference on Big Data Analysis and Deep Learning Applications*. Springer, 2018, pp. 86–93.
8. J. M. Tenenbaum, "Accommodation in computer vision." Stanford Univ Ca Dept of Computer Science, Tech. Rep., 1970.
9. A. Chen, C. Li, S. Zou, M. M. Rahaman, Y. Yao, H. Chen, H. Yang, P. Zhao, W. Hu, W. Liu *et al.*, "Svia dataset: A new dataset of microscopic videos and images for computer-aided sperm analysis," *Biocybernetics and Biomedical Engineering*, 2022.
10. C. Li, H. Chen, X. Li, N. Xu, Z. Hu, D. Xue, S. Qi, H. Ma, L. Zhang, and H. Sun, "A review for cervical histopathology image analysis using machine vision approaches," *Artificial Intelligence Review*, vol. 53, no. 7, pp. 4821–4862, 2020.
11. M. M. Rahaman, C. Li, X. Wu, Y. Yao, Z. Hu, T. Jiang, X. Li, and S. Qi, "A survey for cervical cytopathology image analysis using deep learning," *IEEE Access*, vol. 8, pp. 61 687–61 710, 2020.
12. M. M. Rahaman, C. Li, Y. Yao, F. Kulwa, X. Wu, X. Li, and Q. Wang, "Deepcervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques," *Computers in Biology and Medicine*, vol. 136, p. 104649, 2021.
13. W. Liu, C. Li, M. M. Rahaman, T. Jiang, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun, Y. Yao *et al.*, "Is the aspect ratio of cells important in deep learning? a robust comparison of deep learning methods for multi-scale cytopathology cell image classification: From convolutional neural networks to visual transformers," *Computers in biology and medicine*, p. 105026, 2021.
14. C. Sun, C. Li, J. Zhang, M. M. Rahaman, S. Ai, H. Chen, F. Kulwa, Y. Li, X. Li, and T. Jiang, "Gastric histopathology image segmentation using a hierarchical conditional random field," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1535–1555, 2020.
15. M. M. Rahaman, C. Li, Y. Yao, F. Kulwa, M. A. Rahman, Q. Wang, S. Qi, F. Kong, X. Zhu, and X. Zhao, "Identification of covid-19 samples from chest x-ray images using deep learning: A comparison of transfer learning approaches," *Journal of X-ray Science and Technology*, vol. 28, no. 5, pp. 821–839, 2020.
16. A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, pp. 1–28, 2015.
17. G. M. Dimitri, S. Agrawal, A. Young, J. Donnelly, X. Liu, P. Smielewski, P. Hutchinson, M. Czosnyka, P. Lió, and C. Haubrich, "A multiplex network approach for the analysis of intracranial pressure and heart rate data in traumatic brain injured patients," *Applied network science*, vol. 2, no. 1, pp. 1–12, 2017.
18. V. Cicaloni, O. Spiga, G. M. Dimitri, R. Maiocchi, L. Millucci, D. Giustarini, G. Bernardini, A. Bernini, B. Marzocchi, D. Braconi *et al.*, "Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease," *The FASEB Journal*, vol. 33, no. 11, pp. 12 696–12 703, 2019.
19. A. S. Kwekha-Rashid, H. N. Abduljabbar, and B. Alhayani, "Coronavirus disease (covid-19) cases analysis using machine-learning applications," *Applied Nanoscience*, pp. 1–13, 2021.
20. S. Kosov, K. Shirahama, C. Li, and M. Grzegorzek, "Environmental microorganism classification using conditional random fields and deep convolutional neural networks," *Pattern recognition*, vol. 77, pp. 248–261, 2018.
21. B. Jähne and H. Haußecker, "Computer vision and applications," 2000.
22. C. Li, K. Wang, and N. Xu, "A survey for the applications of content-based microscopic image analysis in microorganism classification domains," *Artificial Intelligence Review*, vol. 51, no. 4, pp. 577–646, 2019.
23. J. Zhang, C. Li, M. M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang, and M. Grzegorzek, "A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches," *Artificial Intelligence Review*, pp. 1–70, 2021.
24. Z. Yan, Y. Zhan, S. Zhang, D. Metaxas, and X. S. Zhou, "Multi-instance multi-stage deep learning for medical image recognition," in *Deep Learning for Medical Image Analysis*. Elsevier, 2017, pp. 83–104.

25. S. Ai, C. Li, X. Li, T. Jiang, M. Grzegorzek, C. Sun, M. M. Rahaman, J. Zhang, Y. Yao, and H. Li, "A state-of-the-art review for gastric histopathology image analysis approaches and future development," *BioMed Research International*, vol. 2021, 2021.
26. H. Chen, C. Li, X. Li, M. M. Rahaman, W. Hu, Y. Li, W. Liu, C. Sun, H. Sun, X. Huang *et al.*, "Il-mcam: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach," *Computers in Biology and Medicine*, p. 105265, 2022.
27. J. Carreira, H. Madeira, and J. G. Silva, "Xception: A technique for the experimental evaluation of dependability in modern computers," *IEEE Transactions on Software Engineering*, vol. 24, no. 2, pp. 125–136, 1998.
28. Q. Guan, Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, and J. Xiang, "Deep convolutional neural network vgg-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," *Journal of Cancer*, vol. 10, no. 20, p. 4876, 2019.
29. A. S. B. Reddy and D. S. Juliet, "Transfer learning with resnet-50 for malaria cell-image classification," in *2019 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2019, pp. 0945–0949.
30. X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2017, pp. 783–787.
31. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
32. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
33. A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1386–1383.
34. S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3634–3642.
35. Z. Chunjiao, "The application and development of photoelectric sensor," in *Intelligence Computation and Evolutionary Computation*. Springer, 2013, pp. 671–677.
36. S. Hata, Y. Saitoh, S. Kumamura, and K. Kaida, "Shape extraction of transparent object using genetic algorithm," in *Proceedings of 13th International Conference on Pattern Recognition*, vol. 4. IEEE, 1996, pp. 684–688.
37. Y. Xu, H. Nagahara, A. Shimada, and R.-i. Taniguchi, "Transcut: Transparent object segmentation from a light-field image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3442–3450.
38. Y. Guo, Z. Xiong, and F. J. Verbeek, "An efficient and robust hybrid method for segmentation of zebrafish objects from bright-field microscope images," *Machine vision and applications*, vol. 29, no. 8, pp. 1211–1225, 2018.
39. A. Nasirahmadi and S.-H. M. Ashtiani, "Bag-of-feature model for sweet and bitter almond classification," *Biosystems engineering*, vol. 156, pp. 51–60, 2017.
40. Y. Xu, K. Maeno, H. Nagahara, A. Shimada, and R.-i. Taniguchi, "Light field distortion feature for transparent object classification," *Computer Vision and Image Understanding*, vol. 139, pp. 122–135, 2015.
41. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
42. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
43. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
44. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
45. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
46. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
47. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
48. C. Li, K. Shirahama, and M. Grzegorzek, "Environmental microbiology aided by content-based image analysis," *Pattern Analysis and Applications*, vol. 19, no. 2, pp. 531–547, 2016.
49. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
50. H. Zhu, H. Jiang, S. Li, H. Li, and Y. Pei, "A novel multispace image reconstruction method for pathological image classification based on structural information," *BioMed research international*, vol. 2019, 2019.
51. H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

52. P. Zhao, C. Li, M. M. Rahaman, H. Xu, P. Ma, H. Yang, H. Sun, T. Jiang, N. Xu, and M. Grzegorzek, "Emds-6: Environmental microorganism image dataset sixth version for image denoising, segmentation, feature extraction, classification, and detection method evaluation," *Frontiers in Microbiology*, p. 1334, 2022.

Resubmited Paper for the Journal： Applied Sciences

Original Paper ID: applsci-1717032

Article Title: A Comparison for Patch-level and Pixel-level Segmentation of Deep Learning Methods on Transparent Environmental Microorganism Images: from Convolutional Neural Networks to Visual Transformers

To: Applied Sciences Editor

Re: Response to reviewers

Dear Editor,

We hope you are well amid this COVID-19 situation.

We express our appreciation for your precious time and effort in reviewing our manuscript and providing us many constructive comments and valuable suggestions. We have now revised our manuscript carefully and thoroughly based on the suggestions and comments. For your convenience, in the revised manuscript, we highlighted the revisions/updates in yellow. We also included our response to the reviewer's comments below. Once again, we greatly appreciate your precious time and effort.

We look forward to your favorable consideration.

We have uploaded three files as follows:

1. Paper (clean, in both LaTeX and PDF);
2. Paper (changes with highlights);
3. Point-by-point response to the comments (below) (response to reviewers).

Best regards,

Chen Li and all authors

Reviewer#1, Comment #1: Abstract section: please perform a review of the section in terms of english editing. there are a few incorrect ways of using capital letters (for example after the comma at line 10). Moreover you should add details on the performances obtained already in the abstract section to give the reader an idea of the results obtained.

Author response: Thank you for your suggestion. Based on your suggestion, we have read the paper carefully and corrected the capital letter errors. At the same time, we have added more details on the performances obtained to the abstract.

Author action: We updated the manuscript by adding in section 1 (Page 1, Line 11-21).

task of transparent images is completed through the reconstruction of pixel patches. In order to facilitate people to understand the performance of different deep learning networks for transparent image segmentation, this paper conducts a series of comparative experiments using patch-level and pixel-level methods. In two sets of experiments, we compared the segmentation performance of four *Convolutional Neural Network* (CNN) models and one *Visual Transformers* (ViT) model on the transparent Environmental Microorganism Data Set Fifth Version dataset, respectively. The research results show that U-Net++ has the highest accuracy rate in the pixel-level comparison experiment with a value of 95.32%. In the patch-level comparison experiment, the highest accuracy rate is ResNet50 with a value of 90.00%. Furthermore, ViT has the lowest accuracy of 89.25% on patch-level segmentation experiments. However, the accuracy rate of ViT in the pixel-level segmentation experiment is 95.31%, second only to U-Nnet++. We conclude that ViT performs the worst in segmentation experiments on pixel-level segmentation, but outperforms most convolutional neural networks on patch-level segmentation. This conclusion is also verifie by the *Environmental Microorganism Data Set Sixth Version dataset* (EMDS-6).

Reviewer#1, Comment #2: Line 31: what does the sentence mean? I think there is the wrong subject. In fact it should be "Deep Learning had good performances..." and then the list of examples. In this list of references I think the authors could expand more, and also add the following relevant references:

-Dimitri, Giovanna Maria, et al. "A multiplex network approach for the analysis of intracranial pressure and heart rate data in traumatic brain injured patients." Applied network science 2.1 (2017): 1-12

-Coudray, Nicolas, et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." Nature medicine 24.10 (2018): 1559-1567.

-Cicaloni, Vittoria, et al. "Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease." The FASEB Journal 33.11 (2019): 12696-12703.

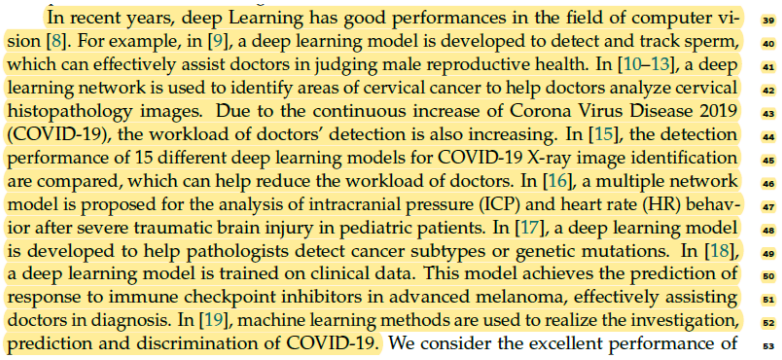
-Tarca, Adi L., et al. "Machine learning and its applications to biology." PLoS computational biology 3.6 (2007): e116.

-Johannet, Paul, et al. "Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma." Clinical Cancer Research 27.1 (2021): 131-140.

-Kwekha-Rashid, Ameer Sardar, Heamn N. Abduljabbar, and Bilal Alhayani. "Coronavirus disease (COVID-19) cases analysis using machine-learning applications." Applied Nanoscience (2021): 1-13.

Author response: Thank you for your suggestion. We've revised the sentence and added relevant references as you suggested.

Author action: We updated the manuscript by adding in section 1 (Page 2, Line 30-53).



In recent years, deep Learning has good performances in the field of computer vision [8]. For example, in [9], a deep learning model is developed to detect and track sperm, which can effectively assist doctors in judging male reproductive health. In [10–13], a deep learning network is used to identify areas of cervical cancer to help doctors analyze cervical histopathology images. Due to the continuous increase of Corona Virus Disease 2019 (COVID-19), the workload of doctors' detection is also increasing. In [15], the detection performance of 15 different deep learning models for COVID-19 X-ray image identification are compared, which can help reduce the workload of doctors. In [16], a multiple network model is proposed for the analysis of intracranial pressure (ICP) and heart rate (HR) behavior after severe traumatic brain injury in pediatric patients. In [17], a deep learning model is developed to help pathologists detect cancer subtypes or genetic mutations. In [18], a deep learning model is trained on clinical data. This model achieves the prediction of response to immune checkpoint inhibitors in advanced melanoma, effectively assisting doctors in diagnosis. In [19], machine learning methods are used to realize the investigation, prediction and discrimination of COVID-19. We consider the excellent performance of

Reviewer#1, Comment #3: Line 46: what does this sentence mean? Please rephrase and expand.

Author response: Thank you for your suggestion. We have supplemented this sentence

as you suggested.

Author action: We updated the manuscript by adding in section 1 (Page 2, Line 64-68).

in images. [24]. For this problem, it is necessary for us to analyze transparent images from patches. We crop the image into fixed-size patches and lead deep learning network to learn the features of the visual information of foreground and background patches. The network trained in this way is sensitive to the foreground and background, which helps to distinguish transparent objects and achieve the purpose of segmentation.

Reviewer#1, Comment #4: Line 52: accumulating convolution layers is not properly correct. Please rephrase.

Author response: Thank you for your suggestion. We reviewed the relevant literature and corrected erroneous descriptions.

Author action: We updated the manuscript by adding in section 1 (Page 2, Line 72-73).

V3 [30], U-Net [31], and novel *Visual Transformers* (ViT) [32]. CNNs gradually expand the receptive field by increasing the size of the convolution kernel until it covers the entire image, so CNNs complete the image extraction from local to global information. In contrast,

Reviewer#1, Comment #5: Line 60 this part concerning the division in test and training is not in the right place. Should be placed in the experimental settings part.

Author response: Thank you for your suggestion. We moved the data division part to the experimental settings part.

Author action: We updated the manuscript by adding in section 3 (Page 3, Line 208-213).

3.1. Experiment Setting

3.1.1. Data Settings

In our work, we use Environmental Microorganism Data Set Fifth Version (EMDS-5) as transparent images for analysis [4]. Tab 1 shows the data distribution of EMDS-5 in the experiment. It is a newly released version of the EMDS series, which includes 21 types of EMs, each of which contains 20 original microscopic images and their corresponding ground truth (GT) images (examples are shown in Fig.3). We randomly divide each category of EMDS-5 into training, validation, and test data sets at a ratio of 1:1:2. Therefore, we

Reviewer#1, Comment #6: Please add more information to Figure 2 caption.

Author response: Thank you for your suggestion. We add more information to Figure 2 caption.

Author action: We updated the manuscript by adding in section 3 (Page 5).

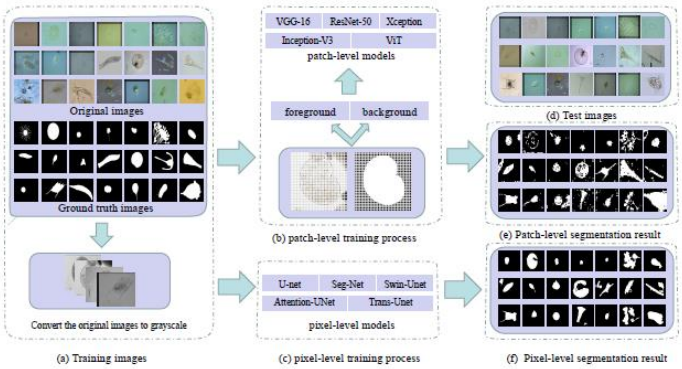


Figure 2. Workflow of patch-level and pixel-level segmentation in transparent images (using environmental microorganism EMDS-5 images as examples) ((a) is the image of the training set and the grayscale of the original image. (b) is the patch-level and pixel-level training process. In (d) is the test set image. (e) and (f) are patch-level and pixel-level segmentation results, respectively).

Reviewer#1, Comment #7: Re-read all of the sections. Qualitative expressions such as "some popular deep learning methods" at line 75-76 should be removed . section 2.1 please re-read and remove qualitative and not scientifically sounding expressions. section 2.2 should change title. Deep Learning is too general. Moreover section 2.3 should be removed.

Author response: Thank you for your suggestion. We have re-read all of the sections and revised the content of the qualitative expression and not scientifically sounding expressions. We removed section 2.3.

Author action: We updated the manuscript by adding in section 2 (Page 3 and 4, Line 88-89, Line 92-95, Line 129-130, Line 135).

2. Related Work

This section briefly introduces the related research on transparent images in practical analysis tasks and the classical deep learning models.

of transparent images of objects (transparent images) is challenging [33]. In the traditional machine learning method, the multi-class fusion algorithm can only extract the shallow features of the transparent image, and the obtain feature layer is incomplete. In practical applications, it is difficult for multi-class fusion algorithms to detect transparent objects.

mentation. The principle is to extract local features of the image for segmentation. However, the foreground transparent objects in transparent images do not have complete features, so these methods are difficult to accurately segment transparent images. The more popular

2.2. Introduction Classic Deep Learning Network Models

Reviewer#1, Comment #8: Table 1: together with number of images it would be usefull to have number of patches obtained later during the pre-processing steps.

Author response: Thank you for your suggestion. We have added tables describing the number of training, validation and test patches.

Author action: We updated the manuscript by adding in section 3 (Page 7).

Table 2. Patch-Level Data Preprocessing. FG (foreground) and BG (background)

Data Set	Training Set	Validation Set	test Set
8×8 pixels FG	16554	17356	32445
8×8 pixels BG	90966	90164	182595
Augmentation With FG	90966	\	\
8×8 Total	181932	107520	215040

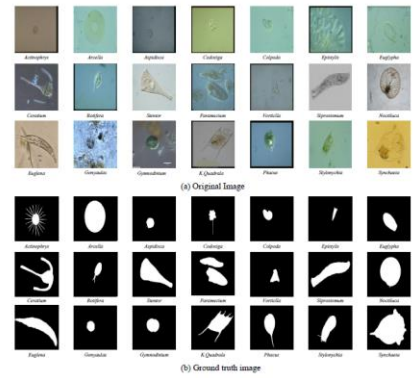
Reviewer#1, Comment #9: In the whole paper there is a tendency of using past and then present tenses with no consistency. Re-read and update accordingly.

Author response: Thank you for your suggestion. We have read the whole paper carefully and updated it to the present tense uniformly.

Reviewer#1, Comment #10: Figure 3 would need more description in the caption.

Author response: Thank you for your suggestion. We have added a more detailed description in the caption of Figure 3.

Author action: We updated the manuscript by adding in section 3 (Page 7).



(a) Original Image

(b) Ground truth image

Figure 3. Examples of the environmental microorganism image in EMDS-5. (a) is the original images of EMDS-5, each image contains one or more EM objects of the same species, and one image is selected for each species as a representative. (b) correspond to the real segmentation images of microorganisms in each image in the (a). The pixel value of the background part in the microorganism image is set to 1, and the foreground part is set to 0.)

Reviewer#1, Comment #11: Line 229 describe more formally the environment on which the experiments has been performed.

Author response: Thank you for your suggestion. We have added a more formal description of the experimental environment.

Author action: We updated the manuscript by adding in section 3 (Page 8, Line 238-242).

3.1.3. Experimental Environment

This comparative experiment is conducted on a local computer. The running memory of the computer is 16 GB. The computer uses Win10 Professional operating system, and it is equipped with an 8 GB NVIDIA Quadro RTX 4000 GPU. In the patch-level experiment, The four CNN network models are imported from keras version 2.3.1 and use tensorflow 2.0.0 as the background. The experimental frameworks for ViT and pixel-level are Pytorch 1.7.1 and Torchvision 8.0.2.

Reviewer#1, Comment #12: Line 250 specify how much the performances improve.

Author response: Thank you for your suggestion. We have added a more detailed description of the performance of the ViT model with and without pretrained weights.

Author action: We updated the manuscript by adding in section 3 (Page 8, Line 264-270).

output space. Finally, the class probability result is output through softmax. Meanwhile, we compare the validation set accuracy of the ViT model with and without pretrained weights. In both sets of experiments, we train three times and then take the average. We find that ViT without pretrained weights and ViT with imagenet pretrained weights had an accuracy of 0.8923 and 0.8926 on the validation set, respectively. During training, ViT takes about 2G less memory than loading the imagenet pre-trained weight model. To compare the performance of the two, we use ViT without pretraining as the optimization option. We

Reviewer#1, Comment #13: Figure 5 add more details to te caption. Figure 6: I would remove as it can be described directly in the text.

Author response: Thank you for your suggestion. We have added a more detailed description in the caption of Figure 5 and removed Figure 6.

Author action: We updated the manuscript by adding in section 3 (Page 9).

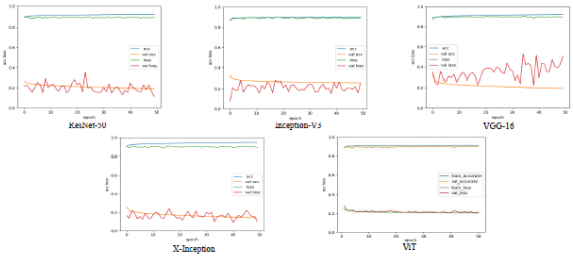


Figure 4. A comparison of the image segmentation results of the loss and accuracy curves of deep learning on 8 × 8 pixels training and validation sets.(Each legend has four curves, respectively, the accuracy and loss values of the training set, and the accuracy and loss values of the validation set.)

Reviewer#1, Comment #14: How were the number of epochs set?

Author response: Thank you for your suggestion. We have added a detailed rationale for the experimental epoch setting.

Author action: We updated the manuscript by adding in section 3 (Page 8, Line 248-261, Line 273-275).

of the loss curve. In our pre-test, we tried to train 100 epochs and keep the best training model weights, and we found that the best model appear between 40 and 50, where too much training caused overfitting and too little training were not able to train the optimal model. Therefore, considering the computational performance of the workstation, we finally set 50 epochs for training. Meanwhile, because of the outstanding classification

sets the batch size to 4. Fig. 5 show the loss curves of different deep learning models in this experiment. We find that the training curves began to converge after 90 epochs of iterations for the five models. To prevent overfitting, we finally set 100 epochs for training.

Reviewer#1, Comment #15: Is it possible to compare your models also using other benchmarks datasets?

Author response: Thank you for your suggestion. We added the *Environmental Microorganism Data Set Sixth Version dataset* (EMDS-6) to compare my model and got the same performance results. The EMDS-6 dataset is a public dataset that has been published in *Frontiers in Microbiology, section Systems Microbiology*. It has the same characteristics as EDMS-5. We summarize the experimental results in the paper.

Author action: We updated the manuscript by adding in section 3 (Page 8 and 9 , Line 370-379).

3.3.3. Additional experiment based on EMDS-6 dataset

To demonstrate the applicability of the models in our comparative experiments, we further compare five models in the pixel-level experiments on the EMDS-6 dataset [52]. The performance metrics of the experimental results of the five models are presented in Tab 8. We find that on EMDS-6, the pixel-level segmentation performance is basically consistent with the segmentation performance on EMDS-5. The segmentation performance of U-Net, U-Net++ and Swin-Unet models is similar, and the segmentation performance of segnet is the worst. In addition, the number of images in EMDS-6 is twice that of EMDS-5, so the model will learn more EMs information during training, which leads to an overall improvement in the segmentation performance of the five models.

Table 8. Segmentation performance of models of five-fold cross-validation experiment on the EMDS-6 test set.

model	Dice	Jaccard	Precision	Recall	Acc
U-Net	84.81	76.24	88.83	83.53	95.43
U-Net++	86.48	78.25	89.02	87.08	95.80
SegNet	74.63	62.50	73.88	83.59	91.21
Trans-Unet	84.66	76.087	86.04	86.88	94.98
Swin-UNet	86.11	78.05	89.46	85.79	95.49

Reviewer#1, Comment #16: Give more indications on the performances given in the 3x3 confusion matrices.

Author response: Thank you for your suggestion. We have added more indications on the performances given in the 3x3 confusion matrices.

Author action: We updated the manuscript by adding in section 3 (Page 11 , Line 320-334).

five models into Fig. 6. We find that the ability of CNNs to classify foreground patches of transparent images is higher than that of ViT. Among them, the best CNN model is Inception-V3, which correctly classify 29686 foreground patches, accounting for 91.50% of the total correct foreground patches. ViT correctly classify 27177 foreground patches, accounting for 83.76% of the total correct foreground patches. In addition, the number of correctly classify backgrounds in ResNet50 is at most 165369, accounting for 90.57% of the total correct background patches, and the Pre of the classify background patches is 97.55%. Among the five models, ResNet50 has the highest prediction accuracy rate of 90.06%. The classification accuracy of X-Inception and Inception models is lower among the five models at 85.85% and 86.30%. Moreover, we find that the X-Inception and Inception models have poor background recognition performance, but better foreground recognition performance. The Inception-V3 model correctly classified up to 29,688 foreground patches, accounting for 91.50% of the total foreground patches. The X-Inception model misclassified a maximum of 27,409 background patches, accounting for 15.01% of the total background patches. The classification performance of the VGG model is relatively moderate among the five models.

Reviewer#1, Comment #17: Figure 10 is too similar to Figure 8 so more details should be given in the captions to let the reader understand differences and similarities between the two.

Author response: Thank you for your suggestion. We have added the caption descriptions of Figure 8 and Figure 10 to make it easier for readers to distinguish .

Author action: We updated the manuscript by adding in section 3 (Page 11 , Line 320-334).

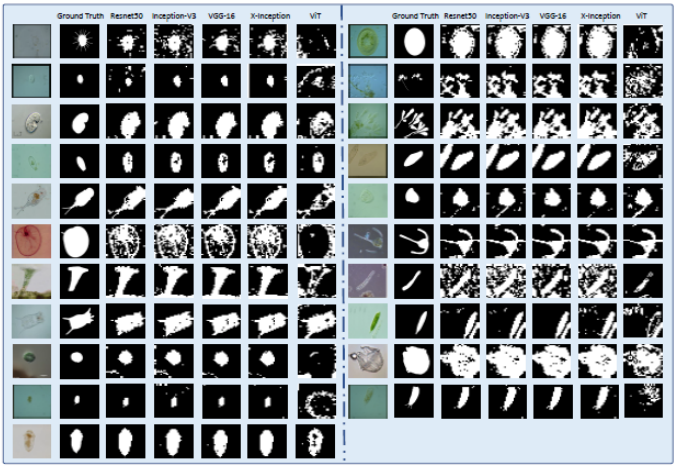


Figure 7. Reconstruct the 8 × 8 pixel patch transparent image segmentation results. (The figure contains the original image, ground truth image and Resnet50, Inception-V3, VGG-16, X-Inception, ViT network model predicted segmentation results.)

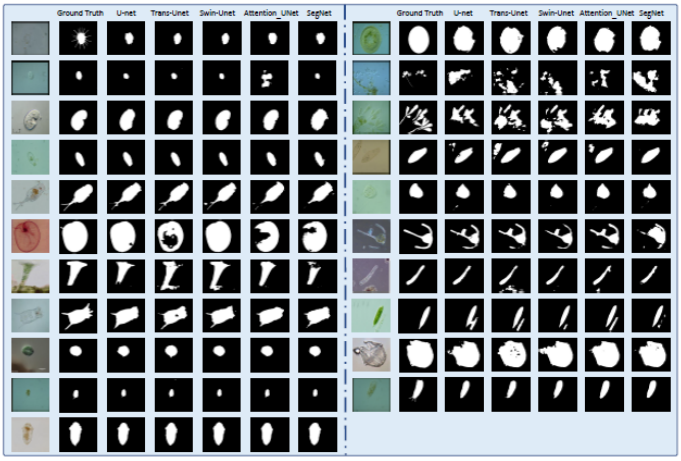


Figure 8. Reconstruction of pixel-level segmentation results on transparent images of the test set. (The figure contains the original image, ground truth image and U-net, Trans-Unet, Swin-Unet, Attention_Unet, SegNet. network model predicted segmentation results.)

Reviewer#1, Comment #18: Figure 5 is not clear.

Author response: Thank you for your suggestion. We have redrawn Figure 5 for the convenience of readers.

Author action: We updated the manuscript by adding in section 3 (Page 11 , Line 320-334).

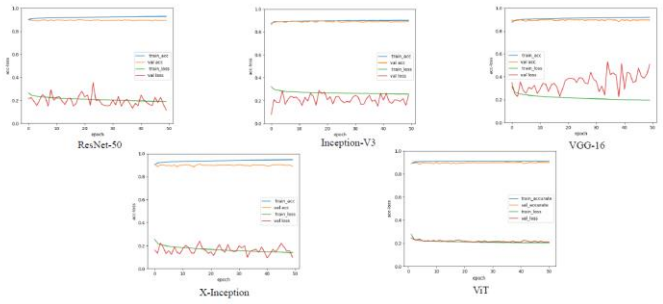


Figure 4. A comparison of the image segmentation results of the loss and accuracy curves of deep learning on 8 × 8 pixels training and validation sets.(Each legend has four curves, respectively, the accuracy and loss values of the training set, and the accuracy and loss values of the validation set.)

Reviewer#1, Comment #19: Did you perform cross validation? Should be explicitly defined.

Author response: Thank you for your suggestion. All our experiments are five-fold cross-validation, and the obtained experimental data are the average of five experiments. We added this detail to the paper.

Author action: We updated the manuscript by adding in section 3 (Page 11 , Line 320-334).

3.3. Comparative Experiment

To avoid network model generalization, we perform five-fold cross-validation in all experiments in this paper. We take the average of the experimentally obtained model performance indicators as the data for the final evaluation model (Precision, Recall, F1-Score, Accuracy, Time, Size, Dice, Jaccard).

Table 4. Classification performance of models of five-fold cross-validation experiment on validation set of 8 × 8 pixels patches. MAcc (Max Acc), FG (foreground) and BG (background) (In [%].)

Model	Class	Avg.Pre	Avg.Rec	Avg.Spe	Avg.F1	MAcc
ResNet50	FG	62.3	88.2	89.7	73.0	92.87
	BG	97.5	89.7	88.2	93.4	
Inception-V3	FG	61.8	88.6	89.5	72.8	90.24
	BG	97.6	89.5	88.6	93.4	
VGG-16	FG	63.1	88.6	90.0	73.7	92.09
	BG	97.6	90.0	88.6	93.6	
X-Inception	FG	53.3	89.2	85.0	66.7	91.10
	BG	96.7	85.0	89.2	90.9	
ViT	FG	62.4	84.1	90.3	71.6	89.26
	BG	96.7	90.3	84.1	93.4	

Table 5. Classification performance of models of five-fold cross-validation experiment on test set of 8 × 8 pixels patches. PAcc (prediction accuracy), FG (foreground) and BG (background)(In [%].)

Model	Class	Avg.Pre	Avg.Rec	Avg.Spe	Avg.F1	Avg.PAcc
ResNet50	FG	62.2	87.2	90.6	72.6	90.0
	BG	97.5	90.6	87.2	93.9	
Inception-V3	FG	52.6	91.5	85.4	66.8	86.29
	BG	98.3	85.4	91.5	91.4	
VGG-16	FG	60.7	89.4	89.7	72.6	89.6
	BG	97.9	89.7	89.4	93.6	
X-Inception	FG	51.8	90.7	85.0	65.9	85.85
	BG	98.1	85.0	90.7	91.1	
ViT	FG	60.4	83.8	90.2	70.2	89.25
	BG	96.9	90.2	83.8	93.4	

Table 6. A comparison of the classification results of five-fold cross-validationexperoment on train and test sets of 8 × 8 pixels patches. Train (Average training time), Test (Average test times) and Avg.p (Single picture prediction time)(In [s].)

model	Train	Test	Avg.p	Size(MB)
ResNet50	36754	878	0.0041	114
Inception-V3	24064	583	0.0027	107
VGG-16	34736	781	0.0036	62.2
X-Inception	46383	1014	0.0047	103
ViT	13992	1308	0.0061	31.2

Table 7. Segmentation performance of models of five-fold cross-validation experiment on the test set

model	Avg.Dice	Avg.Jaccard	Avg.Precision	Avg.Recall	Avg.Acc
U-Net	71.82	59.23	68.98	76.06	91.93
U-Net++	82.51	73.51	83.42	85.98	95.32
SegNet	78.21	67.70	77.45	84.66	74.06
Trans-Unet	75.50	64.13	72.52	86.75	93.44
Swin-UNet	81.00	71.26	85.00	82.08	95.31

Reviewer#2, Comment #1: Never start sentence from “But” - use “However” instead. Line 49 – never start sentence with “And”.

Author response: Thank you for your suggestion. We have read all chapters carefully and corrected this error.

Reviewer#2, Comment #2: Line 3 –“This creastes unnecessary carbon pollution” – I don’t think this sentence is important for the whole article, especially in the abstract

Author response: Thank you for your suggestion. We liked your point very much and

deleted this sentence.

Reviewer#2, Comment #3: Line 6 – “In order …” is incomplete.

Author response: Thank you for your suggestion. We have supplemented this sentence.

Author action: We updated the manuscript by adding in section 3 (Page 1, Line 8-11).

task of transparent images is completed through the reconstruction of pixel patches. In order to facilitate people to understand the performance of different deep learning networks for transparent image segmentation, this paper conducts a series of comparative experiments using patch-level and pixel-level methods. In two sets of experiments, we compared the segmentation performance of four

Reviewer#2, Comment #4: Line 8 – “we crop..” – these are details which are not important in abstract!

Author response: Thank you for your suggestion. We have supplemented this sentence.

Reviewer#2, Comment 5: Line 10 “We” change to “we”.

Author response: I'm sorry for the mistake of capitalization due to my negligence. We have read all sections and corrected them.

Reviewer#2, Comment #6: Line 10 – “ViT” was not explained before – please provide the full name and then use abbreviation. Line 44 – "CNN" was not explained before – it should be done here.

Author response: I am sorry for the mistake of the English abbreviation because of my negligence. We have read all chapters and corrected them.

Reviewer#2, Comment #7: Line 31 “..computer vision had good performance in computer vision acquisition..” – stylistics.

Author response: Thank you for your suggestion. We have supplemented it in detail.

Author action: We updated the manuscript by adding in section 1 (Page 2, Line 30-53).

In recent years, deep Learning has good performances in the field of computer vision [8]. For example, in [9], a deep learning model is developed to detect and track sperm, which can effectively assist doctors in judging male reproductive health. In [10–13], a deep learning network is used to identify areas of cervical cancer to help doctors analyze cervical histopathology images. Due to the continuous increase of Corona Virus Disease 2019 (COVID-19), the workload of doctors' detection is also increasing. In [15], the detection performance of 15 different deep learning models for COVID-19 X-ray image identification are compared, which can help reduce the workload of doctors. In [16], a multiple network model is proposed for the analysis of intracranial pressure (ICP) and heart rate (HR) behavior after severe traumatic brain injury in pediatric patients. In [17], a deep learning model is developed to help pathologists detect cancer subtypes or genetic mutations. In [18], a deep learning model is trained on clinical data. This model achieves the prediction of response to immune checkpoint inhibitors in advanced melanoma, effectively assisting doctors in diagnosis. In [19], machine learning methods are used to realize the investigation, prediction and discrimination of COVID-19. We consider the excellent performance of

Reviewer#2, Comment #8: Line 38 "and contributing to carbon peak and carbon neutrality to a certain extent." – again, it is enough to keep "energy consumption" without mentioning problems of carbonization.

Author response: Thank you for your suggestion. We liked your point very much and deleted this sentence.

Reviewer#2, Comment #9: Line 46 "It is necessary for us to analyze transparent images from patches or pixels. Hence, research work on patch-level and pixel-level are significant for transparent image analysis." – I don't understand what you meant.

Author response: Thank you for your suggestion. We have supplemented this sentence.

Author action: We updated the manuscript by adding in section 1 (Page 2, Line 64-68).

in images. [24]. For this problem, it is necessary for us to analyze transparent images from patches. We crop the image into fixed-size patches and lead deep learning network to learn the features of the visual information of foreground and background patches. The network trained in this way is sensitive to the foreground and background, which helps to distinguish transparent objects and achieve the purpose of segmentation.

Reviewer#2, Comment #10: Line 60 – 65 – this should not be placed in Introduction ! This shell be a part of experiment description. Moreover, Fig. 2 should not be placed so quickly in article. Rather, it should be placed somewhere further.

Author response: Thank you for your suggestion. We moved the partitioning of the dataset and Fig. 2 to the experimental setup subsection.

Author action: We updated the manuscript by adding in section 1 (Page 2 and 6, Line 208-213).

3. Comparative Experiment

This section introduces patch-level and pixel-level segmentation experiments and segmentation results of transparent images under several deep learning networks. The workflow of patch-level and pixel-level image segmentation is shown in Fig. 2.

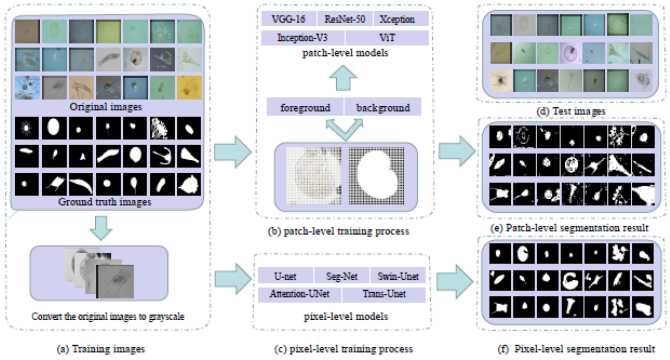


Figure 2. Workflow of patch-level and pixel-level segmentation in transparent images (using environmental microorganism EMDS-5 images as examples) ((a) is the image of the training set and the grayscale of the original image. (b) is the patch-level and pixel-level training process. In (d) is the test set image. (e) and (f) are patch-level and pixel-level segmentation results, respectively).

3.1. Experiment Setting

3.1.1. Data Settings

In our work, we use Environmental Microorganism Data Set Fifth Version (EMDS-5) as transparent images for analysis [4]. Tab 1 shows the data distribution of EMDS-5 in the experiment. It is a newly released version of the EMDS series, which includes 21 types of EMs, each of which contains 20 original microscopic images and their corresponding ground truth (GT) images (examples are shown in Fig.3). We randomly divide each category of EMDS-5 into training, validation, and test data sets at a ratio of 1:1:2. Therefore, we have 105 original images and their corresponding GT images for training and validation, respectively, and 210 original images for testing.

Reviewer#2, Comment #11: Line 100 : just another example of poor English "foreground and background of transparent images are too similar to make analysis difficult." Line 101 – "Compared with deep learning methods, the general traditional analysis methods are time-consuming, labor intensive, and costly." – definitely neural networks can be more power consuming, as they require energy-consuming GPUs to work quickly. Without them, they are very slow on CPUs.

Author response: Thank you for your suggestion. We have removed section 2.3.

Reviewer#3, Comment #1: At the end of the Introduction section (prior to the content organization paragraph), it is necessary to explicitly mention the contribution(s) of the article. It should be clear what the added value of the existing state of the art is.

Author response: Thank you for your suggestion. We fully agree with you and describe the contribution(s) of the article in detail.

Author action: We updated the manuscript by adding in section 1 (Page 2, Line 81-86).

The main contributions of this paper are as follows:
(1) A comparative study on patch-level transparent image segmentation is carried out to help people to analyze transparent images.
(2) The segmentation performance of multiple CNN and ViT deep learning networks under patch-level and pixel-level images are compared, which is convenient for people to do further ensemble learning.

2. Related Work

Reviewer#3, Comment #2: Additionally, make the following adjustments to the document:

- Use the same y-axis range in Figure 5, i.e., between 0 and 1, for the four subplots.
- The title of Figure 6 is not clear. It really corresponds to the complete network, starting from a pre-trained (frozen) model and FC layers for classification.
- Figure 7 should be divided into two figures. One corresponding with the loss curves and the other with the IoU values. It is necessary that the y-axis range is the same between the subfigures.
- Table 2 is not clear, why are there four metrics and three formulas for each? It needs to be rewritten.

Author response: We agree with your suggestion very much, and we have revised the paper according to your suggestion.

Author action: We updated the manuscript by adding in section (Page 2, 10, 14).

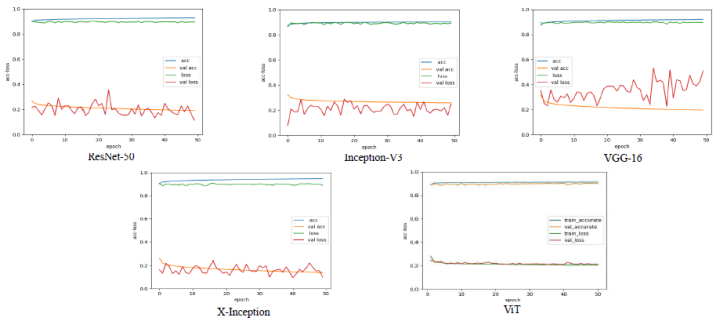


Figure 4. A comparison of the image segmentation results of the loss and accuracy curves of deep learning on 8 × 8 pixels training and validation sets.(Each legend has four curves, respectively, the accuracy and loss values of the training set, and the accuracy and loss values of the validation set.)

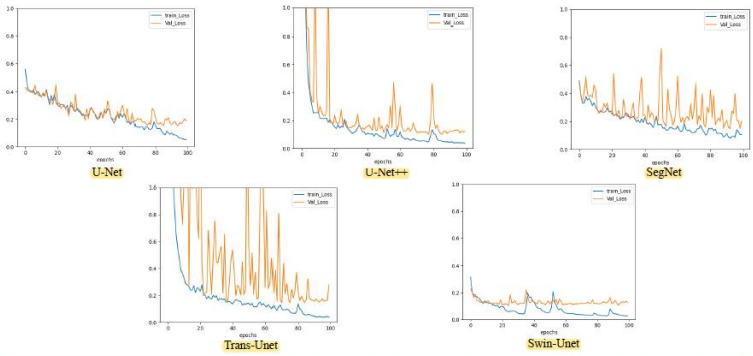


Figure 5. A comparison of the image segmentation results of the loss curves of deep learning on pixel-level training and validation sets.

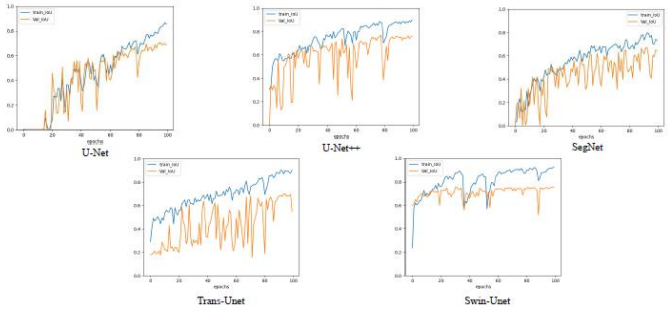


Figure 9. A comparison of the image segmentation results of the IOU curves of deep learning on pixel-level training and validation sets.

Table 3. Evaluation metrics.

Metrics	Formula	Metrics	Formula
Acc	$\frac{TP+TN}{TP+TN+FP+FN}$	Dice	$\frac{2\times\lvert V_{pred}\cap V_{gt}\rvert}{\lvert V_{pred}\rvert+\lvert V_{gt}\rvert}$
Pre (P)	$\frac{TP}{TP+FP}$	Jaccard	$\frac{\lvert V_{pred}\cap V_{gt}\rvert}{\lvert V_{pred}\cup V_{gt}\rvert}$
Rec (R)	$\frac{TP}{TP+FN}$	F1	$\frac{2\times P\times R}{P+R}$
Spe	$\frac{TN}{TN+FP}$		

model pretrained by ImageNet [49] [50]. It is proved that the use of CNN pretrained on ImageNet is useful for classification tasks through the concept of transfer learning and fine-tuning in [51]. Before fine-tuning the pretrained CNN, we freeze the parameters of the pretrained model. After that, we use the patch-level data to fine-tune the dense layers of CNN. We keep the backbone network of the CNN classification network to extract image features, and replace the last fully connected layer of CNN model with Global Average Pooling2D + dense + dense + softmax. Global Average Pooling2D simplifies a large number of parameter operations. The purpose of the dense layer is to extract the correlation between these features through nonlinear changes in the dense layer, and finally map them to the output space. Finally, the class probability result is output through softmax. Meanwhile,

MDPI

Article

A Comparative Study for Patch-level and Pixel-level Segmentation of Deep Learning Methods on Transparent Images of Environmental Microorganisms: from Convolutional Neural Networks to Visual Transformers

Hechen Yang [1], Xin Zhao [1], Tao Jiang [2]*, Jinghua Zhang [1], Peng Zhao [1], Ao Chen [1], Marcin Grzegorzek [3], Shouliang Qi [1], Yueyang Teng [1],Chen Li [1]*

[1] Microscopic Image and Medical Image Analysis Group, MBIE College, Northeastern University, 110169, Shenyang, PR China;
[2] School of Control Engineering, Chengdu University of Information Technology, Chengdu 610225, China;
[3] Institute of Medical Informatics, University of Luebeck, Luebeck, Germany;
* Correspondence: jiang@cuit.edu.cn(T.J.); lichen201096@hotmail.com.(C.L.)

Abstract: Nowadays, the field of transparent image analysis has gradually become a hot topic. However the traditional analysis methods are accompanied by a large number of carbon emissions, and this process consumes a lot of manpower and material resources. With the continuous development of computer vision, it is more appropriate to use computers to analyze images. However, the low contrast between the foreground and background of transparent images makes it difficult for computers to segment transparent objects. For this problem, we start the analysis with pixel patches in the image, and then classify the patches as foreground and background. Finally, the segmentation task of transparent images is completed through the reconstruction of pixel patches. In order to facilitate people to understand the performance of different deep learning networks for transparent image segmentation, this paper conducts a series of comparative experiments using patch-level and pixel-level methods. In two sets of experiments, we compared the segmentation performance of four *Convolutional Neural Network* (CNN) models and one *Visual Transformers* (ViT) model on the transparent Environmental Microorganism Data Set Fifth Version dataset, respectively. The research results show that U-Net++ has the highest accuracy rate in the pixel-level comparison experiment with a value of 95.32%. In the patch-level comparison experiment, the highest accuracy rate is ResNet50 with a value of 90.00%. Furthermore, ViT has the lowest accuracy of 89.25% on patch-level segmentation experiments. However, the accuracy rate of ViT in the pixel-level segmentation experiment is 95.31%, second only to U-Nnet++. We conclude that ViT performs the worst in segmentation experiments on pixel-level segmentation, but outperforms most convolutional neural networks on patch-level segmentation. This conclusion is also verifie by the *Environmental Microorganism Data Set Sixth Version dataset* (EMDS-6).

Keywords: Patch-Level; Pixel-Leve; Image Classification; Image Segmentation; Transparent Images; Deep Learning; Convolutional Neural Network; Visual Transformer;Environmental Microorganism

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

With the advent of the era of science and technology, the application of transparent images has become more and more widely used in various fields around humans, such as the segmentation of renal transparent cancer cell nuclei in medicine [1]. The shape and location information of the cell nucleus is of great significance for the segmentation and diagnosis of benign and malignant renal cancer [2] [3]. Another example is identifying the number of transparent microorganisms in the environment to judge the degree of environmental pollution [4]. In recent years, the segmentation of transparent objects in images is also a hot spot in vision research [5] [6]. It is not an easy task to detect

whether there are transparent objects or translucent objects in images [7]. Because the transparent object area to be observed is generally very small or thin, the colors and contrast of foreground and background are similar. Only the residual edge part leads to the low resolution of foreground or background, which largely depends on its background and lighting conditions. Therefore, there is an urgent need for effective methods to identify transparent or translucent images.

In recent years, deep Learning has good performances in the field of computer vision [8]. For example, in [9], a deep learning model is developed to detect and track sperm, which can effectively assist doctors in judging male reproductive health. In [10–13], a deep learning network is used to identify areas of cervical cancer to help doctors analyze cervical histopathology images. Due to the continuous increase of Corona Virus Disease 2019 (COVID-19), the workload of doctors' detection is also increasing. In [15], the detection performance of 15 different deep learning models for COVID-19 X-ray image identification are compared, which can help reduce the workload of doctors. In [16], a multiple network model is proposed for the analysis of intracranial pressure (ICP) and heart rate (HR) behavior after severe traumatic brain injury in pediatric patients. In [17], a deep learning model is developed to help pathologists detect cancer subtypes or genetic mutations. In [18], a deep learning model is trained on clinical data. This model achieves the prediction of response to immune checkpoint inhibitors in advanced melanoma, effectively assisting doctors in diagnosis. In [19], machine learning methods are used to realize the investigation, prediction and discrimination of COVID-19. We consider the excellent performance of computer vision in image analysis [20], such as high speed, high accuracy, low consumption, high degree of quantification, strong objectivity [21]. In addition, computer analysis of image is more energy-saving and emission-reducing than traditional methods, greatly reducing energy consumption. Therefore computer vision can make up the shortcomings of traditional morphological methods [22]. It brings new opportunities to transparent image analysis [23]. Especially when the object is transparent or has low contrast in the image, we need more foreground information, so we usually find more visual details to recover the lost information from patches or pixels. As shown in Fig 1, the foreground and background of microorganisms are similar. There is only a small amount of information on the edges, so it is difficult for traditional CNN algorithms to globally distinguish transparent objects in images. [24]. For this problem, it is necessary for us to analyze transparent images from patches. We crop the image into fixed-size patches and lead deep learning network to learn the features of the visual information of foreground and background patches. The network trained in this way is sensitive to the foreground and background, which helps to distinguish transparent objects and achieve the purpose of segmentation.

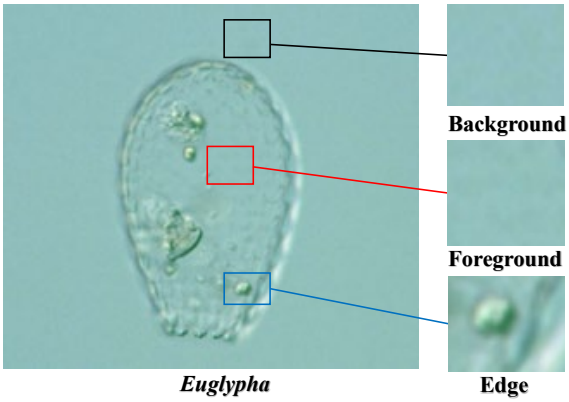


Figure 1. An example of transparent images (a low contrast environmental microorganism image).

In recent years, machine vision has been widely used in image processing [25] [26]. Deep learning is a more effective method in the field of machine vision, such as the popular *Convolutional Neural Network* (CNN) X-Inception [27], VGG-16 [28], Resnet50 [29], Inception-

V3 [30], U-Net [31], and novel *Visual Transformers* (ViT) [32]. CNNs gradually expand the receptive field by increasing the size of the convolution kernel until it covers the entire image, so CNNs complete the image extraction from local to global information. In contrast, transformers can obtain global information from the beginning, making learning more challenging, but their ability to retain long-term dependence is more potent than CNN [32]. Hence, CNNs and Transformers have advantages and disadvantages in dealing with visual information. Therefore, this paper compares patch-level and pixel-level segmentation performance of transparent images with different CNN and Visual Transformers methods. It aims to discover the adaptability of various deep learning models in this research domain.

The main contributions of this paper are as follows:

(1) A comparative study on patch-level transparent image segmentation is carried out to help people to analyze transparent images.

(2) The segmentation performance of multiple CNN and ViT deep learning networks under patch-level and pixel-level images are compared, which is convenient for people to do further ensemble learning.

2. Related Work

This section briefly introduces the related research on transparent images in practical analysis tasks and the classical deep learning models.

2.1. Introduction to Transparent Image Analysis

Object analysis is one of the essential branches in robot vision, especially the analysis of transparent images of objects (transparent images) is challenging [33]. In the traditional machine learning method, the multi-class fusion algorithm can only extract the shallow features of the transparent image, and the obtain feature layer is incomplete. In practical applications, it is difficult for multi-class fusion algorithms to detect transparent objects. For example, home robots can not see things at all when they are detecting some transparent glassware. The ClearGrasp machine learning algorithm performs well in analyzing transparent objects [34]. It can estimate high-precision data of transparent objects from RGB-D transparent images, thereby improving the accuracy of transparent object detection.

As a necessary technical means for analyzing objects, photoelectric sensors are widely used in industrial automation, mechanization, and intelligence. It uses the properties of light to detect the position and change of the object, but when detecting transparent color objects, the light beam of the traditional diffuse reflection photoelectric sensor penetrates the transparent material, causing the sensor to fail. Diffuse reflection photoelectric sensor adopts a phase-locked loop narrowband filter frequency selection technology, which improves the sensitivity to self-returning light and stability of detecting transparent objects [35].

There are many transparent objects in the industrial field, such as transparent plastics, transparent colloids, and liquid drops. These transparent objects bring much uncertainty to products. If factories want to have high-quality products, sometimes it is essential to analyze these transparent objects and control the shapes of the transparent objects. However, it is a difficult problem to segmentation the shape of transparent objects through morphological methods. For instance, Hata et al. use a genetic algorithm to segmentation the transparent paste drop shape in the industry and obtain good performance [36].

The segmentation of transparent objects is very useful in computer vision applications. However, the foreground of a transparent image is usually similar to its background environment, which leads to the general image segmentation methods in dealing with transparent images in general. The light field image segmentation method can accurately and automatically segment transparent images with a small depth of field difference and improve the accuracy of the segmentation, and it has a small amount of calculation [37]. Hence, it is widely used in the segmentation of transparent images.

The correct segmentation of zebrafish in biology has extensively promote the development of life sciences. However, the zebrafish's transparency makes the edges blurre in the

segmentation. The mean shift algorithm can enhance the color representation in the image and improve the discrimination of the specimen against the background [38]. This method improves the efficiency and accuracy of zebrafish specimen segmentation.

Visual object analyze is vital for robotics and computer vision applications. Commonly use statistical analyze methods such as bag-of-features [39] are often applied to image segmentation. The principle is to extract local features of the image for segmentation. However, the foreground transparent objects in transparent images do not have complete features, so these methods are difficult to accurately segment transparent images. The more popular method is the light field distortion feature [40], which can describe transparent objects without knowing the texture of the scene, thus improving the accuracy of segmentation transparent images.

2.2. Introduction Classic Deep Learning Network Models

Simonyan et al. propose the VGG series of deep learning network models (VGG-Net), of which VGG-16 is the most representative [41]. VGG-Net can imitate a larger receptive field by using multiple 3×3 filters, enhancing nonlinear mapping, reducing parameters, and improving the network to be more judgmental. Meanwhile, VGG-16 continues to deepen the previous VGG-Net, with 13 convolutional layers and three fully connected layers. With the continuous increase of convolution kernel and convolution layer, the nonlinear ability of the model is stronger. VGG-16 can better learn the features in images and achieve good performance in analyzing image classification, segmentation, and detection. Simonyan proves that as the depth of the network increases, it promotes the accuracy of image analysis [41]. Nevertheless, this increase in depth is not without a limit. Excessively increasing the depth of the network will lead to network degradation problems. Therefore, the optimal network depth of VGG-Net is set to 16-19 layers. Moreover, VGG-16 has three fully connected layers, which causes more memory to be occupied, too long training time, and difficulty in tuning parameters.

He et al. propose the ResNet series of networks and add a residual structure in networks to solve the problem of network degradation [42]. The ResNet model introduces a jumpy connection method "shortcut connection". This connection method allows the residual structure to skip some levels that not be fully train in the feature extraction process and increases the model's utilization of feature information during the training process. As the most classical model in the ResNet series, ResNet50 has a 50-layer network structure. This model adopts the highway network structure, which makes the network have strong expression capabilities and acquire more advanced features. Therefore, it is widely used in the field of image analysis. However, the network model is too deep and complicated, so how to judge which layers in the deep network not be thoroughly train and then optimize the network is a complex problem.

Szegedy et al. propose the GoogLeNet network model, which has the advantage of reducing the complexity of the network based on ResNet. They first propose Inception-V1, whose network is 22 layers deep and consists of multiple Inception structures cascade as basic modules. Each Inception module consists of a 1×1, 3×3, 5×5 convolution kernel and a 3×3 maximum pooling, which is similar to the idea of multi-scale and increases the adaptability of the network to different scales [43]. With the continuous improvement of the Inception module, the Inception-V2 network uses two 3×3 convolutions instead of 5×5 convolutions. It increases the BN method, which reduces the amount of calculation and speeds up the training time [44]. The Inception-V3 network introduces the idea of decomposing convolution, splitting a larger two-dimensional convolution into two smaller one-dimensional convolutions, further reducing the amount of calculation [45]. At the same time, Inception-V3 optimizes the Inception module embeds the branch in the branch and improves the model's accuracy.

X-Inception is another improvement after Inception-V3 [46]. It mainly uses depth-wise separable convolution to replace the convolution operation in Inception-V3. The X-Inception model uses deep separable convolution to increase the width of the network,

which improves the accuracy of the classification and improves the ability to learn subtle features. Meanwhile, X-Inception adds a residual mechanism similar to ResNet to significantly improve the speed of convergence during training and the model's accuracy. However, X-Inception is relatively fragmented in the calculation process, which results in a slower iteration speed during training.

U-Net is a convolutional neural network, which is initially used to perform the task of medical image segmentation. The architecture of U-Net is symmetrical. It consists of a contracting path and an expansive path [31]. There are two significant contributions of U-Net. The first is the strong use of data augmentation to solve the problem of insufficient training data. The second is its end-to-end structure, which can help the network retrieve the information from the shallow layers. With outstanding performance, U-Net is widely used in semantic segmentation.

Transformer is a deep neural network based on the self-attention mechanism, enabling the model to be trained in parallel and obtain the global information of the training data. Due to its computational efficiency and scalability is widely used in Natural Language Processing. Recently, Dosovitskiy et al. proposed the Vision Transformer (ViT) model and find that it performs very well on image classification tasks [47]. In the first step of training, the ViT model divides pictures into fixed-size image patches and uses its linear sequence as the input of the transformer model. In the second step, position embeddings are added to the embeddings patches to retain the position information, and then the image features are extracted through the multi-head attention mechanism. Finally, the classification model is trained. ViT breaks through the limitation that CNN model cannot be calculated in parallel, and self-attention can produce a more interpretable model. ViT is suitable for solving image processing tasks, but experiments have proved that large data samples are needed to improve the training effect.

3. Comparative Experiment

This section introduces patch-level and pixel-level segmentation experiments and segmentation results of transparent images under several deep learning networks. The workflow of patch-level and pixel-level image segmentation is shown in Fig. 2.

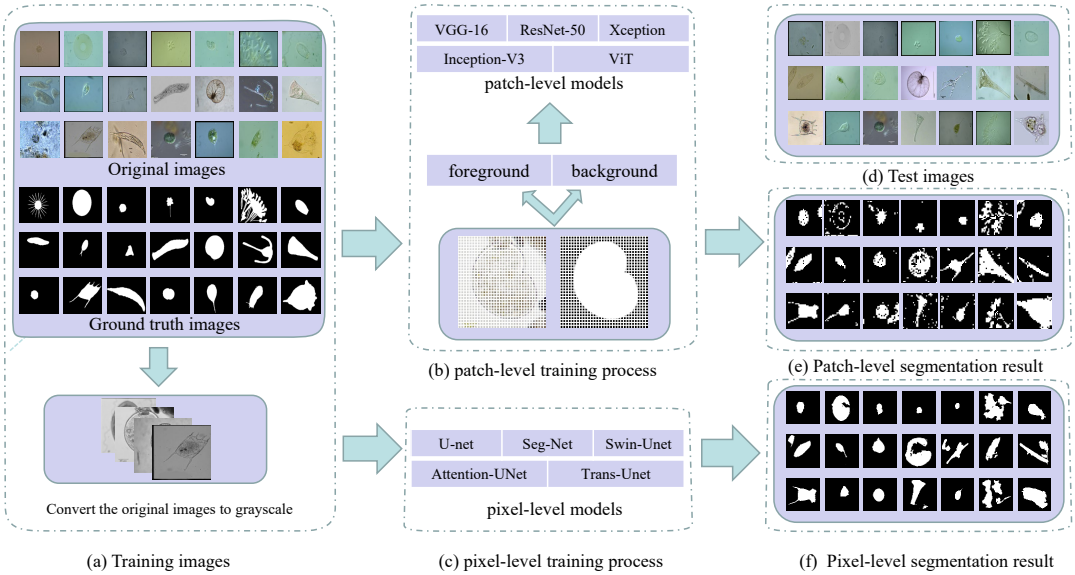


Figure 2. Workflow of patch-level and pixel-level segmentation in transparent images (using environmental microorganism EMDS-5 images as examples) ((a) is the image of the training set and the grayscale of the original image. (b) is the patch-level and pixel-level training process. In (d) is the test set image. (e) and (f) are patch-level and pixel-level segmentation results, respectively).

3.1. Experiment Setting

3.1.1. Data Settings

In our work, we use Environmental Microorganism Data Set Fifth Version (EMDS-5) as transparent images for analysis [4]. Tab 1 shows the data distribution of EMDS-5 in the experiment. It is a newly released version of the EMDS series, which includes 21 types of EMs, each of which contains 20 original microscopic images and their corresponding ground truth (GT) images (examples are shown in Fig.3). We randomly divide each category of EMDS-5 into training, validation, and test data sets at a ratio of 1:1:2. Therefore, we have 105 original images and their corresponding GT images for training and validation, respectively, and 210 original images for testing.

Table 1. EMDS-5 Experimental data.

	Training Set	Validation Set	Test Set
Actinophrys	5	5	10
Arcella	5	5	10
Aspidisca	5	5	10
Codosiga	5	5	10
Colpoda	5	5	10
Epistylis	5	5	10
Euglypha	5	5	10
Paramecium	5	5	10
Rotifera	5	5	10
Vorticlla	5	5	10
Noctiluca	5	5	10
Ceratium	5	5	10
Stentor	5	5	10
Siprostomum	5	5	10
K.Quadrala	5	5	10
Euglena	5	5	10
Gymnodinium	5	5	10
Gonyaulax	5	5	10
Phacus	5	5	10
Stylonychia	5	5	10
Synchaeta	5	5	10
total	105	105	210

3.1.2. Data Preprocessing

Patch-Level Data Preprocessing:

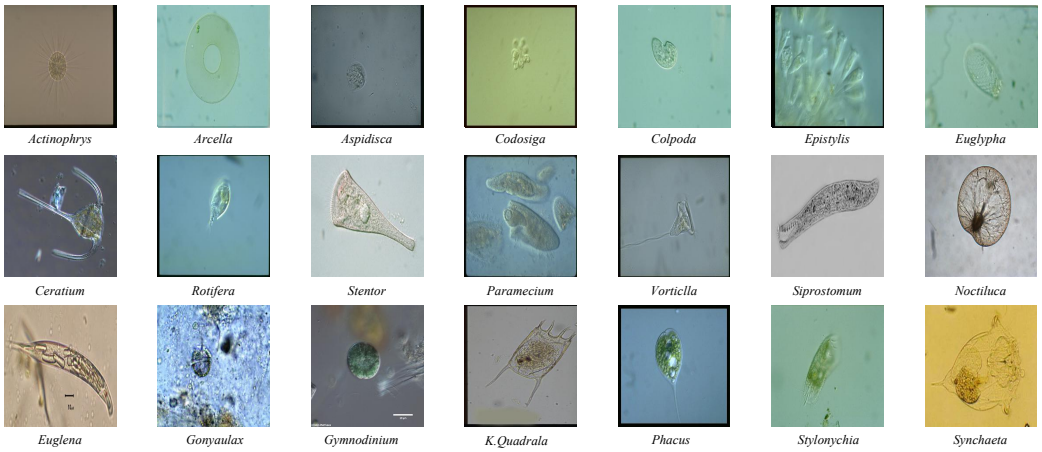
In the first step, considering the colour information is inefficient in EM segmentation [48], these image are converted into grayscale. In the second step, we convert all the image sizes into 256 × 256 pixels uniformly due to the microscopic images having various sizes. In the third step, the training and validation images and their corresponding GT images are cropped into patches (8 × 8 pixels), where 105 × 1024=107520 patches are obtained. We divide these small patches into two categories according to the corresponding GT image small patches: foreground and background. The partition criterion is whether the interest area in the patch takes up half of the whole patch. If it is, we will assign foreground as the label of this patch; otherwise, it is annotated the background. Last step, we find that the 8 × 8 pixels patches with foreground and background are 16554 and 90966, respectively. During the training process, we find that the model weights are heavily biased towards negative samples due to the imbalance of positive and negative samples. In order to avoid data imbalance during training, we rotate the training set image small patches by 0, 90, 180, 270 degrees and mirror them for data augmentation. Then we further obtain 16554×8=132432 patches, from which 90966 patches are randomly selected as the finally used patches in the training set. The details of the image patches are shown in Tab. 2.

Table 2. Patch-Level Data Preprocessing. FG (foreground) and BG (background)

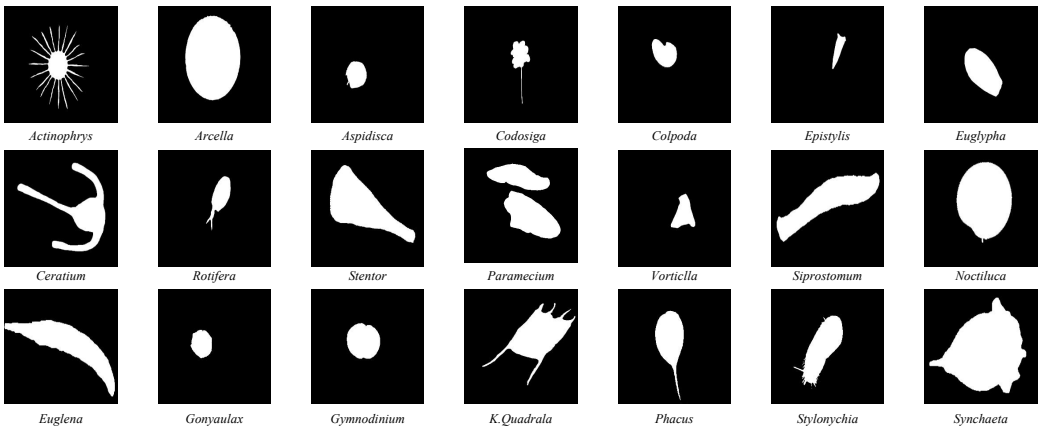
Data Set	Training Set	Validation Set	test Set
8×8 pixels FG	16554	17356	32445
8×8 pixels BG	90966	90164	182595
Augmentation With FG	90966	\	\
8×8 Total	181932	107520	215040

Pixel-Level Data Preprocessing:

We convert the image to grayscale and resize the image size to 256×256 pixel for the pixel-level segmentation experiments.



(a) Original Image



(b) Ground truth image

Figure 3. Examples of the environmental microorganism image in EMDS-5. (a) is the original images of EMDS-5, each image contains one or more EM objects of the same species, and one image is selected for each species as a representative. (b) correspond to the real segmentation images of microorganisms in each image in the (a). The pixel value of the background part in the microorganism image is set to 1, and the foreground part is set to 0.)

3.1.3. Experimental Environment

This comparative experiment is conducted on a local computer. The running memory of the computer is 16 GB. The computer uses Win10 Professional operating system, and it is equipped with an 8 GB NVIDIA Quadro RTX 4000 GPU. In the patch-level experiment, The four CNN network models are imported from keras version 2.3.1 and use tensorflow 2.0.0 as the background. The experimental frameworks for ViT and pixel-level are Pytorch 1.7.1 and Torchvision 8.0.2.

3.1.4. Hyper Parameters

The patch-level experiment uses the Adam optimizer with a 0.0002 learning rate and sets the batch size to 32. In Fig. 4 show the accuracy and loss curves of different deep learning models in this experiment. The Epoch is determined according to the convergence of the loss curve. In our pre-test, we tried to train 100 epochs and keep the best training model weights, and we found that the best model appear between 40 and 50, where too much training caused overfitting and too little training were not able to train the optimal model. Therefore, considering the computational performance of the workstation, we finally set 50 epochs for training. Meanwhile, because of the outstanding classification ability of CNN in ImageNet and the significant performance of transfer learning with limited training data set [41], we use the limited EM training data to fine-tune the CNN model pretrained by ImageNet [49] [50]. It is proved that the use of CNN pretrained on ImageNet is useful for classification tasks through the concept of transfer learning and fine-tuning in [51]. Before fine-tuning the pretrained CNN, we freeze the parameters of the pretrained model. After that, we use the patch-level data to fine-tune the dense layers of CNN. We keep the backbone network of the CNN classification network to extract image features, and replace the last fully connected layer of CNN model with Global Average Pooling2D + dense + dense + softmax. Global Average Pooling2D simplifies a large number of parameter operations. The purpose of the dense layer is to extract the correlation between these features through nonlinear changes in the dense layer, and finally map them to the output space. Finally, the class probability result is output through softmax. Meanwhile, we compare the validation set accuracy of the ViT model with and without pretrained weights. In both sets of experiments, we train three times and then take the average. We find that ViT without pretrained weights and ViT with imagenet pretrained weights had an accuracy of 0.8923 and 0.8926 on the validation set, respectively. During training, ViT takes about 2G less memory than loading the imagenet pre-trained weight model. To compare the performance of the two, we use ViT without pretraining as the optimization option. We set the network depth to 6, heads to 16, mlp_dim to 3000, dropout and emb_dropout to 0.1. The pixel-level experiment uses the Adam optimizer with a 0.001 learning rate and sets the batch size to 4. Fig. 5 show the loss curves of different deep learning models in this experiment. We find that the training curves began to converge after 90 epochs of iterations for the five models. To prevent overfitting, we finally set 100 epochs for training.

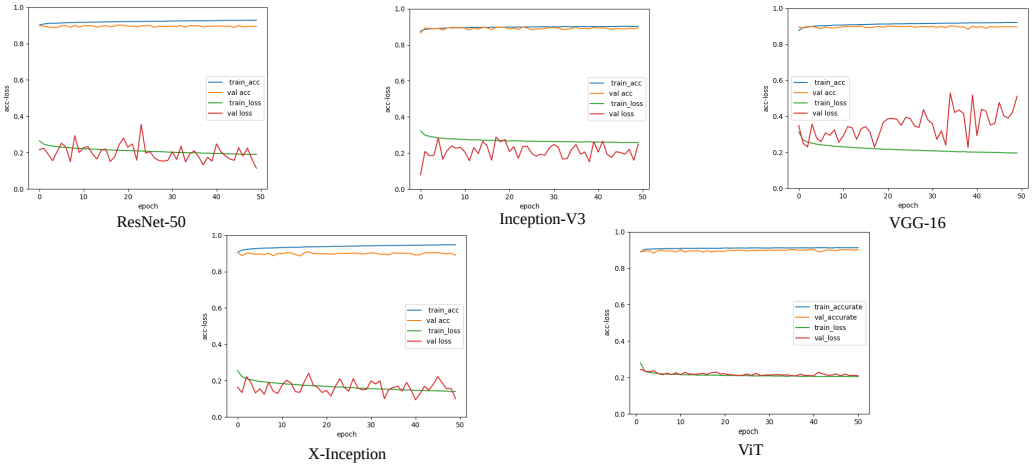


Figure 4. A comparison of the image segmentation results of the loss and accuracy curves of deep learning on 8×8 pixels training and validation sets.(Each legend has four curves, respectively, the accuracy and loss values of the training set, and the accuracy and loss values of the validation set.)

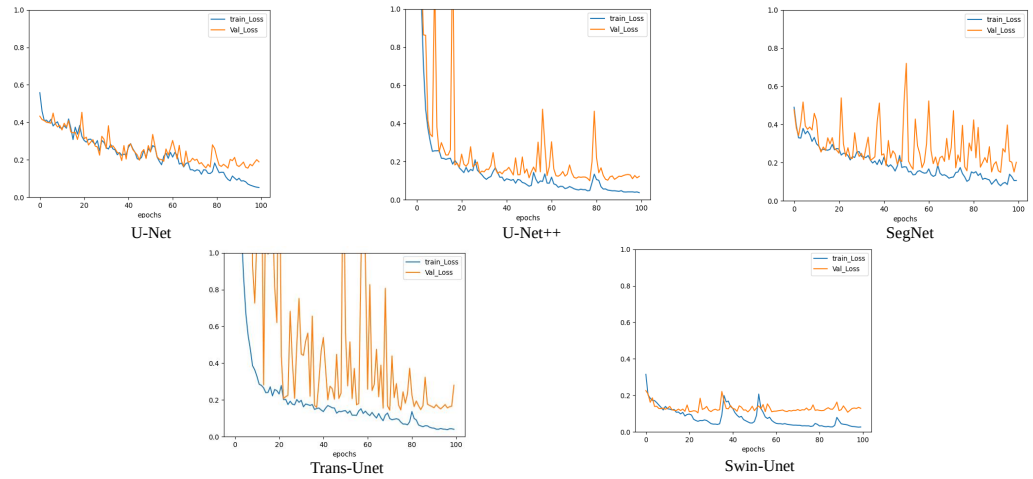


Figure 5. A comparison of the image segmentation results of the loss curves of deep learning on pixel-level training and validation sets.

3.2. Evaluation Metrics

To compare the classification foreground and background performance of different methods, we used the commonly used deep learning classification indexes Accuracy (Acc), Precision (Pre), Recall (Rec), Specificity (Spe), and F1-Score (F1) to evaluate the patch-level results [52]. Acc reflects the ratio of correct classification samples to total samples. Pre reflects the proportion of correctly predicted positive samples in model classification's positive samples. Rec reflects the correct proportion of model classification in whole positive samples. Spe reflects the proportion of the model correctly classifying the negative samples in the total negative samples. F1 is a calculation result that comprehensively considers the Pre and Rec of the model. Besides that, we employ Dice, Jaccard, Pre, Acc, and Rec to evaluate the results in pixel-level segmentation [16]. V_{pred} represents the foreground that the model predicts. V_{gt} represents the foreground in a ground truth image. From Tab 3, we can find that the higher the values of the first four metrics (Dice, Jaccard, Recall, and Accuracy) are, the better the segmentation results are. TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative) are concepts in the confusion matrix.

Table 3. Evaluation metrics.

Metrics	Formula	Metrics	Formula
Acc	$\frac{TP+TN}{TP+TN+FP+FN}$	Dice	$\frac{2\times\lvert V_{pred}\cap V_{gt}\rvert}{\lvert V_{pred}\rvert+\lvert V_{gt}\rvert}$
Pre (P)	$\frac{TP}{TP+FP}$	Jaccard	$\frac{\lvert V_{pred}\cap V_{gt}\rvert}{\lvert V_{pred}\cup V_{gt}\rvert}$
Rec (R)	$\frac{TP}{TP+FN}$	F1	$\frac{2\times P\times R}{P+R}$
Spe	$\frac{TN}{TN+FP}$		

3.3. Comparative Experiment

To avoid network model generalization, we perform five-fold cross-validation in all experiments in this paper. We take the average of the experimentally obtained model performance indicators as the data for the final evaluation model (Precision, Recall, F1-Score, Accuracy, Time, Size, Dice, Jaccard).

3.3.1. Comparative Experiment of Patch-level Segmentation

Comparison on Training and Validation Sets:

In order to compare the classification performance of CNNs and ViT models, we calculate precision, recall, F1-Score, and max accuracy are used to evaluate the models. The 5-class classification results of 8×8 pixels patches on validation set are presented in Tab 4. Overall, the Pre of the deep learning network classifying the transparent image background is higher than the foreground. Besides, the Pre of the five models to classify transparent images backgrounds is almost 97%; the highest is the VGG-16 value of 97.6%, and the lowest is the X-Inception and the ViT value of 96.7%. Meanwhile, the Pre rate of classification foreground VGG-16 is the best, and the Pre rate is 63.1%. The Inception-V3 obtains the lowest 53.3%. For transparent images foreground classification, the highest Rec rate is obtained with X-Inception, which is 89.2%, and the lowest one is ViT, which is 84.1%.For transparent images background classification, the highest Rec rate is the Vit value of 90.3%, and the lowest is the X-Inception value of 85.0%. The Spe obtained by the five models in the classification background is opposite to the Rec rate obtained in the classification foreground. Among the five models, the highest Acc is ResNet50 with a value of 92.87%, and the lowest is ViT with 89.26%.

Table 4. Classification performance of models of five-fold cross-validation experiment on validation set of 8×8 pixels patches. MAcc (Max Acc), FG (foreground) and BG (background) (In [%].)

Model	Class	Avg.Pre	Avg.Rec	Avg.Spe	Avg.F1	MAcc
ResNet50	FG	62.3	88.2	89.7	73.0	92.87
	BG	97.5	89.7	88.2	93.4	
Inception-V3	FG	61.8	88.6	89.5	72.8	90.24
	BG	97.6	89.5	88.6	93.4	
VGG-16	FG	63.1	88.6	90.0	73.7	92.09
	BG	97.6	90.0	88.6	93.6	
X-Inception	FG	53.3	89.2	85.0	66.7	91.10
	BG	96.7	85.0	89.2	90.9	
ViT	FG	62.4	84.1	90.3	71.6	89.26
	BG	96.7	90.3	84.1	93.4	

Comparison on Test Set:

In Tab 5 we summarize the results of these five network predictions. We can find that Acc of ResNet50 is the highest (90.00%), Acc of X-Inception is the lowest at 85.85%. Furthermore, the lowest prediction Acc of the transparent foreground is the X-Inception value of 51.8%, and the highest is the ResNet50 value of 62.2%.

In order to more intuitively express the classification results of CNN and ViT models for transparent image patches, we summarize the confusion matrices predicted by

Table 5. Classification performance of models of five-fold cross-validation experiment on test set of 8×8 pixels patches. PAcc (prediction accuracy), FG (foreground) and BG (background)(In [%].)

Model	Class	Avg.Pre	Avg.Rec	Avg.Spe	Avg.F1	Avg.PAcc
ResNet50	FG	62.2	87.2	90.6	72.6	90.0
	BG	97.5	90.6	87.2	93.9	
Inception-V3	FG	52.6	91.5	85.4	66.8	86.29
	BG	98.3	85.4	91.5	91.4	
VGG-16	FG	60.7	89.4	89.7	72.6	89.6
	BG	97.9	89.7	89.4	93.6	
X-Inception	FG	51.8	90.7	85.0	65.9	85.85
	BG	98.1	85.0	90.7	91.1	
ViT	FG	60.4	83.8	90.2	70.2	89.25
	BG	96.9	90.2	83.8	93.4	

five models into Fig. 6. We find that the ability of CNNs to classify foreground patches of transparent images is higher than that of ViT. Among them, the best CNN model is Inception-V3, which correctly classify 29686 foreground patches, accounting for 91.50% of the total correct foreground patches. ViT correctly classify 27177 foreground patches, accounting for 83.76% of the total correct foreground patches. In addition, the number of correctly classify backgrounds in ResNet50 is at most 165369, accounting for 90.57% of the total correct background patches, and the Pre of the classify background patches is 97.55%. Among the five models, ResNet50 has the highest prediction accuracy rate of 90.06%. The classification accuracy of X-Inception and Inception models is lower among the five models at 85.85% and 86.30%. Moreover, we find that the X-Inception and Inception models have poor background recognition performance, but better foreground recognition performance. The Inception-V3 model correctly classified up to 29,688 foreground patches, accounting for 91.50% of the total foreground patches. The X-Inception model misclassified a maximum of 27,409 background patches, accounting for 15.01% of the total background patches. The classification performance of the VGG model is relatively moderate among the five models. To better show the classification results, we reconstruct the transparent image after dicing in Fig. 7.

In Tab 6, we provide the model training and prediction time and the size of the model during the experiment. From the perspective of model training time, the ViT model is much lower than CNN models, where the ViT training time is 13992 seconds, and the X-Inception training time is the longest, 46383 seconds. From the perspective of the model's size, the minimum size of the ViT model is 31.2M, and the maximum size of the ResNet50 model is 114M. We calculate the time of the five prediction models. The fastest prediction time of Inception-V3 is 583 seconds, and the prediction time of a single picture is 0.0027 seconds. The slowest time of ViT is 1308 seconds, and the prediction time of a single image is 0.0061 seconds.

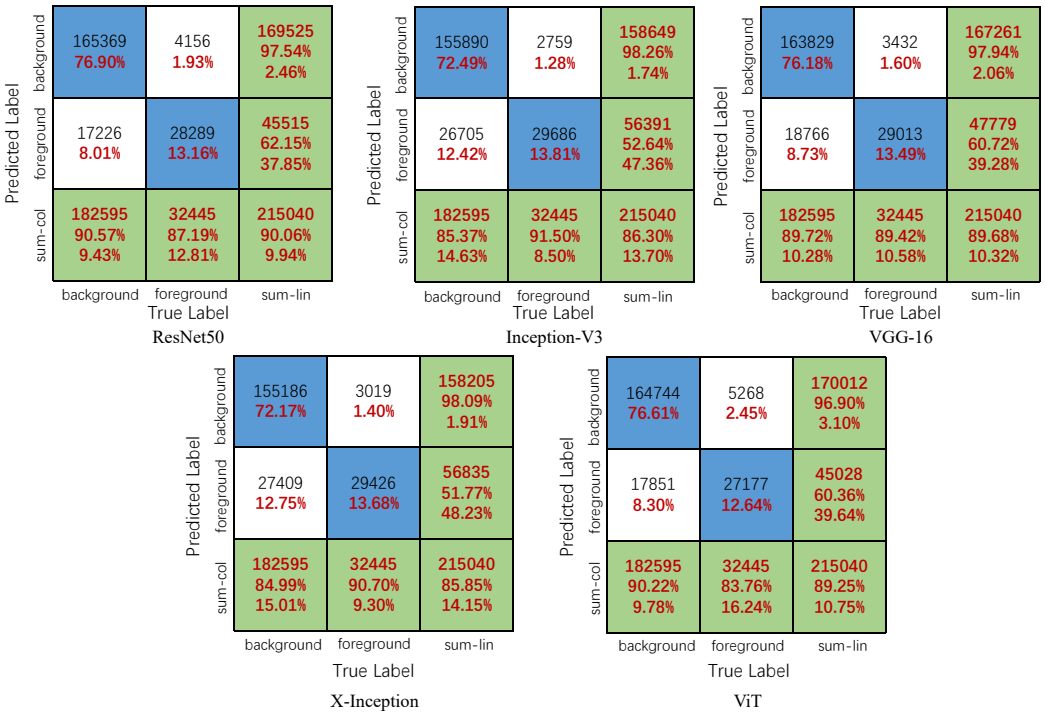


Figure 6. Predict the confusion matrix on test set of 8 × 8 pixels patches

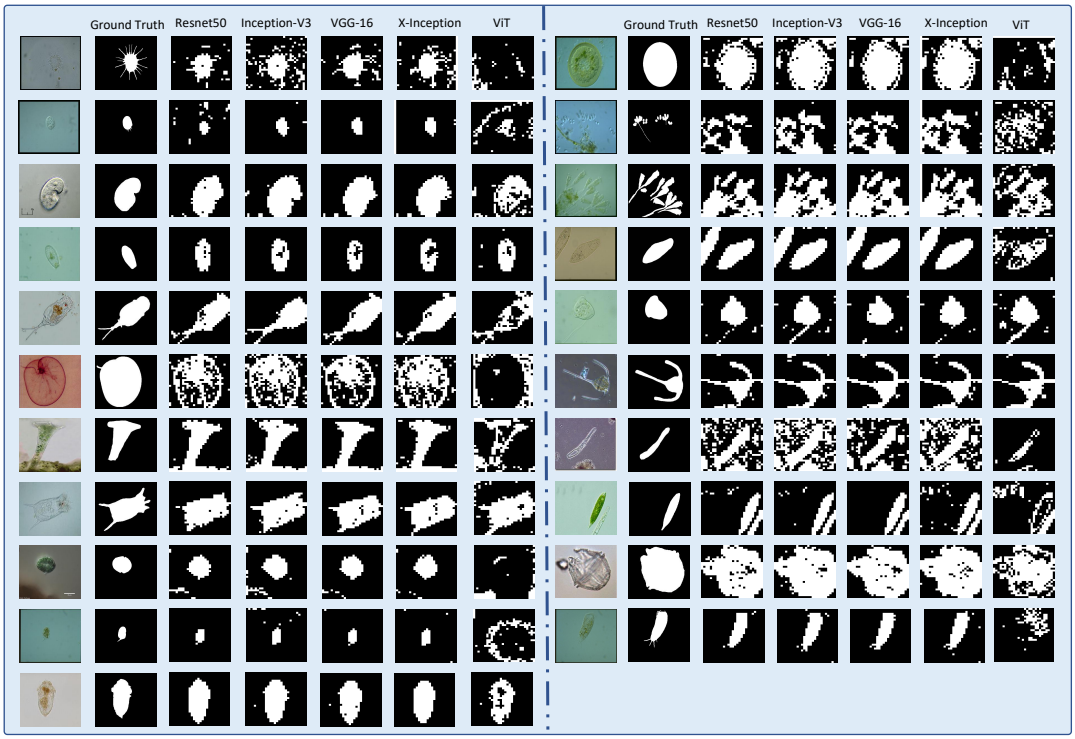


Figure 7. Reconstruct the 8 × 8 pixel patch transparent image segmentation results. (The figure contains the original image, ground truth image and Resnet50, Inception-V3, VGG-16, X-Inception, ViT network model predicted segmentation results.)

Table 6. A comparison of the classification results of five-fold cross-validationexperoment on train and test sets of 8×8 pixels patches. Train (Average training time), Test (Average test times) and Avg.p (Single picture prediction time)(In [s].)

model	Train	Test	Avg.p	Size(MB)
ResNet50	36754	878	0.0041	114
Inception-V3	24064	583	0.0027	107
VGG-16	34736	781	0.0036	62.2
X-Inception	46383	1014	0.0047	103
ViT	13992	1308	0.0061	31.2

3.3.2. Comparison Experiment of Pixel-Level Segmentation

To compare the effect of path-level segmentation, we conduct extended experiments on pixel-level segmentation. We apply five networks for comparative experiments: U-Net, U-Net++, SegNet, TransUnet, and Swin-UNet. We use these five networks to compare the performance of CNN and *visual transformer* (VT) for pixel-level segmentation. U-Net, U-Net++, SegNet stands for CNN network, Swin-UNet stands for transformer networks, TransUnet stands for CNNs joins transformer. In tab 7, we show five model prediction outcome metrics. We find that U-Net++ has the highest segmentation performance in the whole, but it also has the longest training time. U-Net has the worst segmentation performance. The segmentation result of vision transformer network (Swin-UNet) is second only to U-Net++. Its Jaccard and precision values of 71.26% and 85.00%, which are higher than other network models. In order to compare the segmentation results more intuitively, we show the pixel-level segmentation results in Fig. 8. We can see that pixel-level segmentation results are generally better than patch-level. However, the patch-level segmentation effect is better on multi-object transparent microorganism images. Compared with the 8×8 patch-level segmentation, the network model with transformer structure(Swin-UNet) at the pixel level performs well, and the VT is higher than the accuracy of the CNN network model. However, in the 8×8 patch-level experiment, the accuracy of CNN networks are higher than ViT. In Fig. 5, we find that the Loss curve stability of Swin-UNet is significantly better than the other four models during training. The stability of the training loss of the VT model is better than that of the CNN. In order to more intuitively reflect the training process of the model, we show the *Intersection-over-Union* (IOU) curves of the five models on the training set and the validation set in the pixel-level experiment in Fig. 9.

Table 7. Segmentation performance of models of five-fold cross-validation experiment on the test set

model	Avg.Dice	Avg.Jaccard	Avg.Precision	Avg.Recall	Avg.Acc
U-Net	71.82	59.23	68.98	76.06	91.93
U-Net++	82.51	73.51	83.42	85.98	95.32
SegNet	78.21	67.70	77.45	84.66	74.06
Trans-Unet	75.50	64.13	72.52	86.75	93.44
Swin-UNet	81.00	71.26	85.00	82.08	95.31

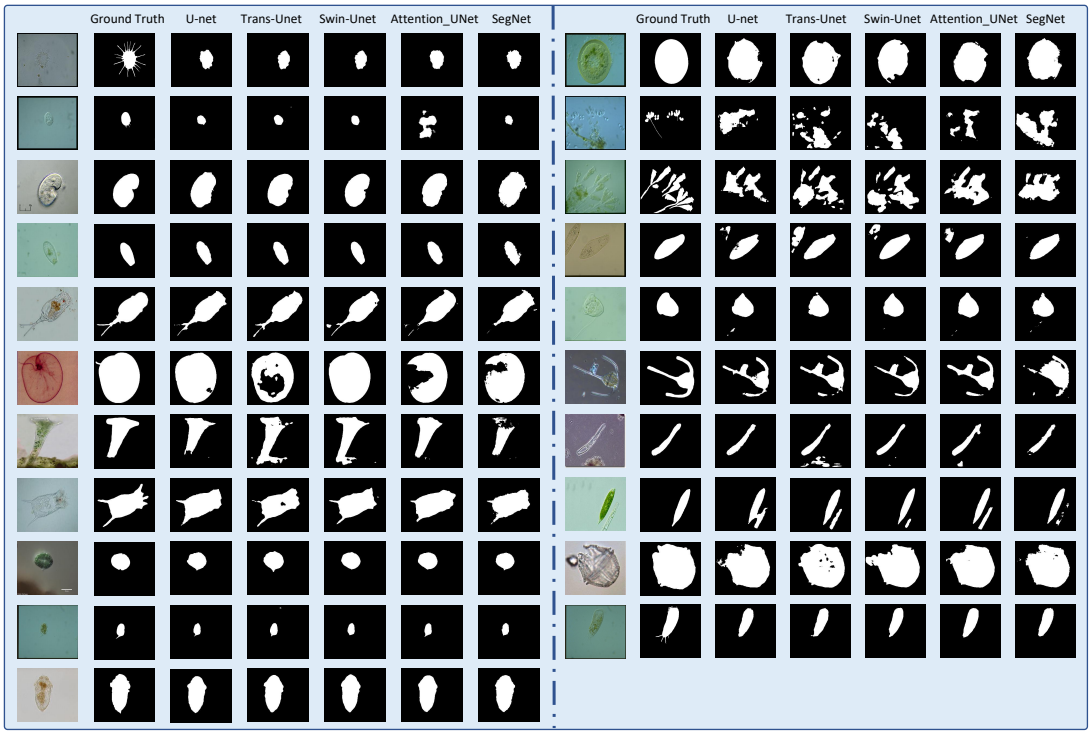


Figure 8. Reconstruction of pixel-level segmentation results on transparent images of the test set. (The figure contains the original image, ground truth image and U-net, Trans-Unet, Swin-Unet, Attention_Unet, SegNet. network model predicted segmentation results.)

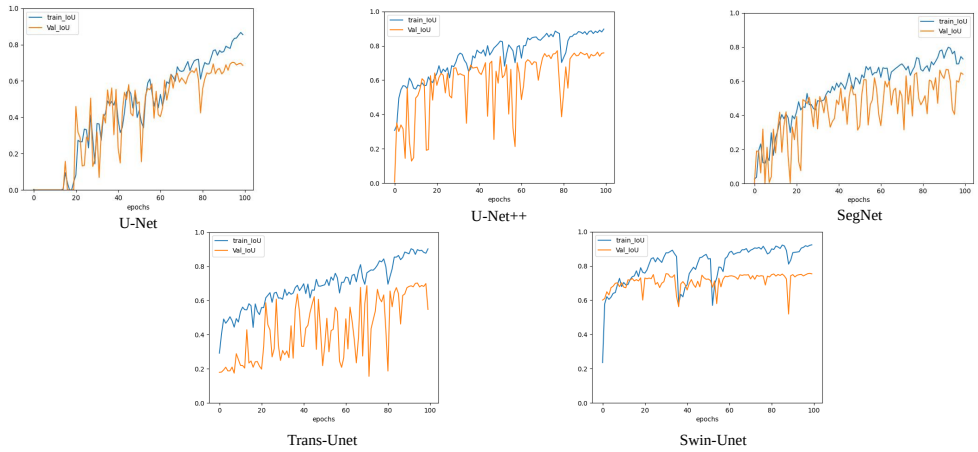


Figure 9. A comparison of the image segmentation results of the IOU curves of deep learning on pixel-level training and validation sets.

3.3.3. Additional experiment based on EMDS-6 dataset

To demonstrate the applicability of the models in our comparative experiments, we further compare five models in the pixel-level experiments on the EMDS-6 dataset [52]. The performance metrics of the experimental results of the five models are presented in Tab 8. We find that on EMDS-6, the pixel-level segmentation performance is basically consistent with the segmentation performance on EMDS-5. The segmentation performance of U-Net, U-Net++ and Swin-Unet models is similar, and the segmentation performance of segnet is the worst. In addition, the number of images in EMDS-6 is twice that of EMDS-5, so the model will learn more EMs information during training, which leads to an overall improvement in the segmentation performance of the five models.

Table 8. Segmentation performance of models of five-fold cross-validation experiment on the EMDS-6 test set.

model	Dice	Jaccard	Precision	Recall	Acc
U-Net	84.81	76.24	88.83	83.53	95.43
U-Net++	86.48	78.25	89.02	87.08	95.80
SegNet	74.63	62.50	73.88	83.59	91.21
Trans-Unet	84.66	76.087	86.04	86.88	94.98
Swin-UNet	86.11	78.05	89.46	85.79	95.49

3.4. In-depth Analysis

In the predicted 215040 patches, we compare the performance of five types of network classification foreground and background. In Fig. 6, we find that Inception-v3 has the largest number of correct foregrounds under 8×8 pixels patches. ResNet50 has the largest number of correctly classify backgrounds. We find that Inception-v3 has the largest number of correctly classify foregrounds under 224×224 pixels patches, and the largest number of correctly classify background patches is ViT. In addition, the number of foreground patches misclassify by the ViT network model is much smaller than that of the CNNs network. At the same time, the number of correctly classify foregrounds in the CNNs network is greater than that of the ViT network.

In the predicted 215040 patches, we compare the performance of five types of network classification foreground and background. In Fig. 6, we find that VGG-16 has the largest number of correct foregrounds under 8×8 pixels patches. Inception-v3 has the largest number of correctly classify backgrounds. However, the number of correctly classified foregrounds of ViT is higher than that of VGG-16, Inception-V3 and X-Inception. Besides, the ability of Swin-UNet to segment foreground also outperforms most models. So, VT model is also outstanding for low-transparency image recognition.

4. Conclusion and Future Work

In this paper, we aim at the problem that transparent images are difficult to segmentation by cropping the image into patches and classifying the foreground and background. We use CNNs and VT deep learning methods to compare patch-level and pixel-level performance of the segmentation of transparent images. We find that pixel-level generally outperforms patch-level in segmenting transparent microorganism images. However patch-level works better in multi-object segmentation. In addition, in the patch-level segmentation experiment, CNNs are better than the VT model, but in the pixel-level experiment, the VT model segmentation performance is better than most CNNs. When the patch pixel is smaller, the more regions perceived by the VT model, the stronger the ability to combine contextual information. In addition, the loss convergence and stability of the VT model during training are better than the CNN model. The VT model has great potential in the future. Therefore, CNN and ViT models have more advantages in image classification. CNN is good at extracting local features of images, while ViT is good at extracting global features of images combined with contextual information.

In the future, we plan to increase the amount of data to improve the stability of the comparison. Meanwhile, the images reconstructed by deep learning classification can be extended to the positioning, recognition, and detection of transparent images. We will further strengthen the application of results.

Author Contributions: Conceptualization, C.L.; methodology, H.Y. and C.L.; software, H.Y.; validation, P.Z., A.C. and H.Y.; formal analysis, H.Y.; investigation, M.G. and T.J.; resources, X.Z.; data curation, C.L. and H.Y.; writing—original draft preparation, H.Y. and C.L.; writing—review and editing, C.L., J.Z. and H.Y.; visualization, H.Y.; supervision, C.L.; project administration, C.L.; funding acquisition, C.L. and T.J. All authors have read and agreed to the published version of the manuscript.

Funding: National Natural Science Foundation of China (No. 61806047).

Acknowledgments: We thank Miss Zixian Li and Mr. Guoxian Li for their important discussion.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. S.-Y. Liao, O. N. Aurelio, K. Jan, J. Zavada, and E. J. Stanbridge, "Identification of the mn/ca9 protein as a reliable diagnostic biomarker of clear cell carcinoma of the kidney," *Cancer research*, vol. 57, no. 14, pp. 2827–2831, 1997.
2. D. Xue, X. Zhou, C. Li, Y. Yao, M. M. Rahaman, J. Zhang, H. Chen, J. Zhang, S. Qi, and H. Sun, "An application of transfer learning and ensemble learning techniques for cervical histopathology image classification," *IEEE Access*, vol. 8, pp. 104 603–104 618, 2020.
3. X. Zhou, C. Li, M. M. Rahaman, Y. Yao, S. Ai, C. Sun, Q. Wang, Y. Zhang, M. Li, X. Li *et al.*, "A comprehensive review for breast histopathology image analysis using classical and deep neural networks," *IEEE Access*, vol. 8, pp. 90 931–90 956, 2020.
4. Z. Li, C. Li, Y. Yao, J. Zhang, M. M. Rahaman, H. Xu, F. Kulwa, B. Lu, X. Zhu, and T. Jiang, "Emds-5: Environmental microorganism image dataset fifth version for multiple image analysis tasks," *Plos one*, vol. 16, no. 5, p. e0250631, 2021.
5. J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, and H. Li, "Lcu-net: A novel low-cost u-net for environmental microorganism image segmentation," *Pattern Recognition*, vol. 115, p. 107885, 2021.
6. F. Kulwa, C. Li, X. Zhao, B. Cai, N. Xu, S. Qi, S. Chen, and Y. Teng, "A state-of-the-art survey for microorganism image segmentation methods and future potential," *IEEE Access*, vol. 7, pp. 100 243–100 269, 2019.
7. M. P. Khaing and M. Masayuki, "Transparent object detection using convolutional neural network," in *International Conference on Big Data Analysis and Deep Learning Applications*. Springer, 2018, pp. 86–93.
8. J. M. Tenenbaum, "Accommodation in computer vision." Stanford Univ Ca Dept of Computer Science, Tech. Rep., 1970.
9. A. Chen, C. Li, S. Zou, M. M. Rahaman, Y. Yao, H. Chen, H. Yang, P. Zhao, W. Hu, W. Liu *et al.*, "Svia dataset: A new dataset of microscopic videos and images for computer-aided sperm analysis," *Biocybernetics and Biomedical Engineering*, 2022.
10. C. Li, H. Chen, X. Li, N. Xu, Z. Hu, D. Xue, S. Qi, H. Ma, L. Zhang, and H. Sun, "A review for cervical histopathology image analysis using machine vision approaches," *Artificial Intelligence Review*, vol. 53, no. 7, pp. 4821–4862, 2020.
11. M. M. Rahaman, C. Li, X. Wu, Y. Yao, Z. Hu, T. Jiang, X. Li, and S. Qi, "A survey for cervical cytopathology image analysis using deep learning," *IEEE Access*, vol. 8, pp. 61 687–61 710, 2020.
12. M. M. Rahaman, C. Li, Y. Yao, F. Kulwa, X. Wu, X. Li, and Q. Wang, "Deepcervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques," *Computers in Biology and Medicine*, vol. 136, p. 104649, 2021.
13. W. Liu, C. Li, M. M. Rahaman, T. Jiang, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun, Y. Yao *et al.*, "Is the aspect ratio of cells important in deep learning? a robust comparison of deep learning methods for multi-scale cytopathology cell image classification: From convolutional neural networks to visual transformers," *Computers in biology and medicine*, p. 105026, 2021.
14. C. Sun, C. Li, J. Zhang, M. M. Rahaman, S. Ai, H. Chen, F. Kulwa, Y. Li, X. Li, and T. Jiang, "Gastric histopathology image segmentation using a hierarchical conditional random field," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1535–1555, 2020.
15. M. M. Rahaman, C. Li, Y. Yao, F. Kulwa, M. A. Rahman, Q. Wang, S. Qi, F. Kong, X. Zhu, and X. Zhao, "Identification of covid-19 samples from chest x-ray images using deep learning: A comparison of transfer learning approaches," *Journal of X-ray Science and Technology*, vol. 28, no. 5, pp. 821–839, 2020.
16. A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, pp. 1–28, 2015.
17. G. M. Dimitri, S. Agrawal, A. Young, J. Donnelly, X. Liu, P. Smielewski, P. Hutchinson, M. Czosnyka, P. Lió, and C. Haubrich, "A multiplex network approach for the analysis of intracranial pressure and heart rate data in traumatic brain injured patients," *Applied network science*, vol. 2, no. 1, pp. 1–12, 2017.
18. V. Cicaloni, O. Spiga, G. M. Dimitri, R. Maiocchi, L. Millucci, D. Giustarini, G. Bernardini, A. Bernini, B. Marzocchi, D. Braconi *et al.*, "Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease," *The FASEB Journal*, vol. 33, no. 11, pp. 12 696–12 703, 2019.
19. A. S. Kwekha-Rashid, H. N. Abduljabbar, and B. Alhayani, "Coronavirus disease (covid-19) cases analysis using machine-learning applications," *Applied Nanoscience*, pp. 1–13, 2021.
20. S. Kosov, K. Shirahama, C. Li, and M. Grzegorzek, "Environmental microorganism classification using conditional random fields and deep convolutional neural networks," *Pattern recognition*, vol. 77, pp. 248–261, 2018.
21. B. Jähne and H. Haußecker, "Computer vision and applications," 2000.
22. C. Li, K. Wang, and N. Xu, "A survey for the applications of content-based microscopic image analysis in microorganism classification domains," *Artificial Intelligence Review*, vol. 51, no. 4, pp. 577–646, 2019.
23. J. Zhang, C. Li, M. M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang, and M. Grzegorzek, "A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches," *Artificial Intelligence Review*, pp. 1–70, 2021.
24. Z. Yan, Y. Zhan, S. Zhang, D. Metaxas, and X. S. Zhou, "Multi-instance multi-stage deep learning for medical image recognition," in *Deep Learning for Medical Image Analysis*. Elsevier, 2017, pp. 83–104.

25. S. Ai, C. Li, X. Li, T. Jiang, M. Grzegorzek, C. Sun, M. M. Rahaman, J. Zhang, Y. Yao, and H. Li, "A state-of-the-art review for gastric histopathology image analysis approaches and future development," *BioMed Research International*, vol. 2021, 2021.
26. H. Chen, C. Li, X. Li, M. M. Rahaman, W. Hu, Y. Li, W. Liu, C. Sun, H. Sun, X. Huang *et al.*, "Il-mcam: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach," *Computers in Biology and Medicine*, p. 105265, 2022.
27. J. Carreira, H. Madeira, and J. G. Silva, "Xception: A technique for the experimental evaluation of dependability in modern computers," *IEEE Transactions on Software Engineering*, vol. 24, no. 2, pp. 125–136, 1998.
28. Q. Guan, Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, and J. Xiang, "Deep convolutional neural network vgg-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," *Journal of Cancer*, vol. 10, no. 20, p. 4876, 2019.
29. A. S. B. Reddy and D. S. Juliet, "Transfer learning with resnet-50 for malaria cell-image classification," in *2019 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2019, pp. 0945–0949.
30. X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2017, pp. 783–787.
31. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
32. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
33. A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1386–1383.
34. S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3634–3642.
35. Z. Chunjiao, "The application and development of photoelectric sensor," in *Intelligence Computation and Evolutionary Computation*. Springer, 2013, pp. 671–677.
36. S. Hata, Y. Saitoh, S. Kumamura, and K. Kaida, "Shape extraction of transparent object using genetic algorithm," in *Proceedings of 13th International Conference on Pattern Recognition*, vol. 4. IEEE, 1996, pp. 684–688.
37. Y. Xu, H. Nagahara, A. Shimada, and R.-i. Taniguchi, "Transcut: Transparent object segmentation from a light-field image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3442–3450.
38. Y. Guo, Z. Xiong, and F. J. Verbeek, "An efficient and robust hybrid method for segmentation of zebrafish objects from bright-field microscope images," *Machine vision and applications*, vol. 29, no. 8, pp. 1211–1225, 2018.
39. A. Nasirahmadi and S.-H. M. Ashtiani, "Bag-of-feature model for sweet and bitter almond classification," *Biosystems engineering*, vol. 156, pp. 51–60, 2017.
40. Y. Xu, K. Maeno, H. Nagahara, A. Shimada, and R.-i. Taniguchi, "Light field distortion feature for transparent object classification," *Computer Vision and Image Understanding*, vol. 139, pp. 122–135, 2015.
41. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
42. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
43. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
44. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
45. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
46. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
47. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
48. C. Li, K. Shirahama, and M. Grzegorzek, "Environmental microbiology aided by content-based image analysis," *Pattern Analysis and Applications*, vol. 19, no. 2, pp. 531–547, 2016.
49. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
50. H. Zhu, H. Jiang, S. Li, H. Li, and Y. Pei, "A novel multispace image reconstruction method for pathological image classification based on structural information," *BioMed research international*, vol. 2019, 2019.
51. H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

52. P. Zhao, C. Li, M. M. Rahaman, H. Xu, P. Ma, H. Yang, H. Sun, T. Jiang, N. Xu, and M. Grzegorzek, "Emds-6: Environmental microorganism image dataset sixth version for image denoising, segmentation, feature extraction, classification, and detection method evaluation," *Frontiers in Microbiology*, p. 1334, 2022.